Docket No. 826.1610/JDH

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Koji TSUKAMOTO

Serial No.: Not Assigned

Filed: June 14, 2000

For: APPARATUS FOR RETRIEVING INFORMATION USING REFERENCE REASON OF DOCUMENT

Group Art Unit: Not Assigned

Examiner: Not Assigned

#2

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. §1.55

Honorable Commissioner of
Patents and Trademarks
Washington, D.C. 2023l

Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicant(s) submit(s) herewith a certified copy of the following foreign application:

Japanese Patent Application No. 11-168552, filed: June 15, 1999.

It is respectfully requested that the applicants be given the benefit of the foreign filing date as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. §119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: June 14, 2000

By: _____

James D. Halsey, Jr.
Registration No. 22,729

700 Eleventh Street, N.W., Suite 500
Washington, D.C. 20001
(202) 434-1500

PATENT OFFICE
JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of the following application as filed with this Office.

Date of Application: June 15, 1999

Application Number: Patent Application No. 11-168552

Applicant(s): FUJITSU LIMITED

April 28, 2000

Commissioner,
Patent Office Takahiko KONDO

Certificate No.2000-3031664

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９９年　６月１５日

出 願 番 号
Application Number: 平成１１年特許願第１６８５５２号

出 願 人
Applicant (s): 富士通株式会社

CERTIFIED COPY OF
PRIORITY DOCUMENT

２０００年　４月２８日

特 許 庁 長 官
Commissioner,
Patent Office 近 藤 隆 彦

出証番号　出証特２０００－３０３１６６４

【書類名】　特許願

【整理番号】　9900604

【提出日】　平成11年 6月15日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 17/27

G06F 17/30

【発明の名称】　文書の参照理由を用いて情報検索を行う装置

【請求項の数】　15

【発明者】

【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

【氏名】　塚本　浩司

【特許出願人】

【識別番号】　000005223

【氏名又は名称】　富士通株式会社

【代理人】

【識別番号】　100074099

【住所又は居所】　東京都千代田区二番町８番地２０　二番町ビル３Ｆ

【弁理士】

【氏名又は名称】　大菅　義之

【電話番号】　03-3238-0031

【選任した代理人】

【識別番号】　100067987

【住所又は居所】　神奈川県横浜市港北区太尾町１４１８－３０５（大倉山二番館）

【弁理士】

【氏名又は名称】　久木元　彰

【電話番号】　045-545-9280

【手数料の表示】

【予納台帳番号】　012542

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9705047

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書の参照理由を用いて情報検索を行う装置

【特許請求の範囲】

【請求項１】　与えられた文書データから被参照文書に関する文書情報を抽出する被参照文書抽出手段と、

前記文書データ内で前記被参照文書を参照している位置に関する情報を抽出する参照位置抽出手段と、

前記参照位置抽出手段により抽出された情報を解析して、前記被参照文書が参照されている理由を判断する判断手段と、

前記被参照文書抽出手段により抽出された情報と前記被参照文書が参照されている理由とを含む出力情報を出力する出力手段と

を備えることを特徴とする参照理由同定装置。

【請求項２】　前記文書データから該文書データに関する文書情報を抽出する文書情報抽出手段をさらに備え、前記出力手段は、該文書情報抽出手段により抽出された情報を前記出力情報とともに出力することを特徴とする請求項１記載の参照理由同定装置。

【請求項３】　前記判断手段は、前記被参照文書を参照している位置が属する章の情報と、該被参照文書を参照している位置の周辺の文字列の情報のうち少なくとも一方に基づいて、前記被参照文書が参照されている理由を判断することを特徴とする請求項１記載の参照理由同定装置。

【請求項４】　与えられた文書データから被参照文書に関する文書情報を抽出する被参照文書抽出手段と、

前記文書データ内で前記被参照文書を参照している位置に関する情報を抽出する参照位置抽出手段と、

前記参照位置抽出手段により抽出された情報を解析して、前記被参照文書が参照されている理由を判断する判断手段と、

前記被参照文書を参照している位置の周辺の情報から、前記被参照文書のためのキーワード情報を抽出するキーワード抽出手段と、

前記被参照文書抽出手段により抽出された情報と前記被参照文書が参照されて

いる理由と前記キーワード情報とを含む出力情報を出力する出力手段と
を備えることを特徴とするキーワード抽出装置。

【請求項５】　文書データを格納する文書データベース手段と、
ある文書が被参照文書を参照している理由を表す参照理由を含む参照関係を格納する参照関係格納手段と、
前記参照関係格納手段に格納された参照関係を利用して、前記文書データベース手段に格納された文書データの検索を行う検索手段と、
前記参照理由を含む検索結果を出力する出力手段と
を備えることを特徴とする情報検索装置。

【請求項６】　前記ある文書の文書データから前記被参照文書に関する文書情報を抽出する被参照文書抽出手段と、
前記ある文書の文書データ内で前記被参照文書を参照している位置に関する情報を抽出する参照位置抽出手段と、
前記参照位置抽出手段により抽出された情報を解析して、前記参照理由を判断する判断手段と、
前記被参照文書抽出手段により抽出された情報と前記参照理由とを含む情報を前記参照関係格納手段に格納する手段と
をさらに備えることを特徴とする請求項５記載の情報検索装置。

【請求項７】　前記参照関係格納手段は、前記ある文書から抽出された前記被参照文書のためのキーワード情報を格納し、前記検索手段は、該キーワード情報を利用して検索を行うことを特徴とする請求項５記載の情報検索装置。

【請求項８】　前記出力手段は、前記参照関係を前記参照理由に基づいてグラフィカルに表示する表示手段を含むことを特徴とする請求項５記載の情報検索装置。

【請求項９】　前記出力手段は、前記参照関係を時系列に表示する表示手段を含むことを特徴とする請求項５記載の情報検索装置。

【請求項１０】　与えられた文書データから被参照文書に関する文書情報を抽出する被参照文書抽出手段と、
前記文書データ内で前記被参照文書を参照している位置に関する情報を抽出す

る参照位置抽出手段と、

前記参照位置抽出手段により抽出された情報を解析して、前記被参照文書が参照されている理由を判断する判断手段と、

複数の文書データ間において、被参照文書が参照されている理由を含む参照関係の類似度を計算し、該複数の文書データを分類する類似度判定手段と、

分類結果を出力する出力手段と

を備えることを特徴とする文書分類装置。

【請求項１１】　前記被参照文書を参照している位置の周辺の情報から、前記被参照文書のためのキーワード情報を抽出するキーワード抽出手段をさらに備え、前記類似度判定手段は、該キーワード情報を用いて前記複数の文書データを分類することを特徴とする請求項１０記載の文書分類装置。

【請求項１２】　検索対象の情報を格納するデータベース手段と、

ある情報が被参照情報を参照している理由を表す参照理由を含む参照関係を格納する参照関係格納手段と、

前記参照関係格納手段に格納された参照関係を利用して、前記データベース手段に格納された情報の検索を行う検索手段と、

前記参照理由を含む検索結果を出力する出力手段と

を備えることを特徴とする情報検索装置。

【請求項１３】　コンピュータのためのプログラムを記録した記録媒体であって、

与えられた文書データから被参照文書に関する文書情報を抽出するステップと、

前記文書データ内で前記被参照文書を参照している位置に関する情報を抽出するステップと、

前記被参照文書を参照している位置に関する情報を解析して、前記被参照文書が参照されている理由を判断するステップと、

前記被参照文書に関する文書情報と前記被参照文書が参照されている理由とを含む出力情報を出力するステップと

を含む処理を前記コンピュータに実行させるためのプログラムを記録したコン

ピュータ読み取り可能な記録媒体。

【請求項１４】　コンピュータのためのプログラムを記録した記録媒体であって、

ある文書が被参照文書を参照している理由を表す参照理由を含む参照関係を利用して、文書データベースに格納された文書データの検索を行うステップと、

前記参照理由を含む検索結果を出力するステップと

を含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項１５】　コンピュータのためのプログラムを記録した記録媒体であって、

与えられた文書データから被参照文書に関する文書情報を抽出するステップと、

前記文書データ内で前記被参照文書を参照している位置に関する情報を抽出するステップと、

前記被参照文書を参照している位置に関する情報を解析して、前記被参照文書が参照されている理由を判断する判断ステップと、

複数の文書データ間において、被参照文書が参照されている理由を含む参照関係の類似度を計算し、該複数の文書データを分類するステップと、

分類結果を出力するステップと

を含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】

本発明は、文書等の情報を検索する情報検索に係り、特に、文書間に参照／被参照関係が存在する場合に、参照理由を同定し、参照理由を用いて文書を検索する装置に関する。

【０００２】

【従来の技術】

文書間に参照／被参照関係が存在する場合に、参照関係を検索に利用する技術がいくつか提案されている。このような技術としては、以下のような公開特許公報が挙げられる。

（１）特開昭６３－２２８２２１（三菱電機）

文献の参照関係を記憶しておき、手がかりとなる何らかの情報からファジイ論理演算を用いて検索を行う。

（２）特開昭６３－１５３６３０（日本電気）

文献の被引用関係を用いて、引用文献を検索項目とすることにより、共引用関係の文書の検索を行うことができる。ここで、共引用関係の文書とは、共通の引用文献を持つことで重要な関係にあると考えられる文献を意味する。

（３）特開平１－１９１２５８（リコー）

文献の本文と、本文中から自動的に抽出した参考文献名とを同時に提示して、編集を容易にする。

（４）特開平６－２８２５３４（日本電気）

文書が引用されたことを、引用された文書の利用者に自動的に通知する。

（５）特開平７－３１１７８０（キャノン）

引用関係に基づき、ある文献に関連する文献を検索する。検索結果は、重要度順に表示される。重要度は、引用回数に基づいて決められる。

（６）特開平８－２７２８１８（新日本製鐵）

ある文書を指定すると、それに関連する文書が表示され、表示された文書を選択することにより、さらに検索が可能になる。

（７）特開平９－１４６９６８（日立製作所）

ある文献が行っている参照と似たような参照を行っている他の文献を検索する。

（８）特開平１０－１０５５７２（日本電気）

参照関係とキーワードをもとに、文書間に関連性があるかないかを判断し、それに基づいて文書集合を作成する。

【０００３】

図２７は、このような従来の情報検索システムを示している。このシステムは

、検索装置１、全文データベース２、および参照関係データベース３を備える。そして、ある文献が入力されると、それと共引用関係にある文献や似たような参照を行っている文献を、関連文献として表示する。検索装置１の検索部５は、全文データベース２の文献を検索し、選択部６は、参照関係データベース３の参照関係を用いて関連文献を選択する。

【０００４】

関連文献の情報は、例えば、図２８のような形式で表示される。関連文献が複数ある場合、例えば、文献の重要度の順に並べて表示される。

上述の公開特許公報のうち、特開平８－２７２８１８と特開平１０－１０５５７２以外においては、基本的に、参照しているかいないかのどちらかという２値的な参照関係のみが用いられている。

【０００５】

また、特開平８－２７２８１８では、表示の際にどのような位置付けの参照関係なのかを表示するようにしているが、そのためには人間があらかじめ情報を付与しておく必要がある。また、特開平１０－１０５５７２では、参照関係とキーワードを用いて、文書間の関連性があるかないかを判断している。

【０００６】

このように、従来の技術では、関連があるかないかのいずれか一方を表す２値情報が、文書間の関連性として用いられている。

【０００７】

【発明が解決しようとする課題】

しかしながら、上述した従来の情報検索システムには、次のような問題がある。

【０００８】

特開平７－３１１７８０では、引用回数の多い文献から順に重要文献として表示しているが、利用者は装置側が一意に提示した価値基準に従って文献の重要度を逐次判断せざるを得ない。

【０００９】

また、他の従来技術も、文書を参照しているかいないかの２値情報で参照関係

を表しており、ある文書にとって「重要であるかないか」という１つの観点から分析を行い、文書を検索するものである。このような２値情報を用いて検索を行うシステムは、次のような欠点を持っている。

（１）利用者になぜ重要であるのかを示すことができない。

（２）重要であるかないかという１つの観点に基づく分析では、文書同士の関わり合いを示すことができない。

（３）どのような類の文書が必要であるかが分かっていても、すべての関連文書が一緒に表示されてしまう。

【００１０】

このようなシステムで、例えば、科学分野の文献を検索した場合、新しい技術の文献や、よく用いられるソフトウェアについての文献というような、主題として直接的な関連がないものがしばしば上位に表示されることがある。実際に参考にしたい文献はこのような文献ではない場合が多いが、従来のシステムでは、この問題を避けることができない。

【００１１】

本発明の課題は、ある文書がどのような理由で被参照文書を参照しているのかを同定し、同定された参照理由を利用して効率的に文書を検索する装置を提供することである。

【００１２】

【課題を解決するための手段】

図１は、本発明の装置の原理図である。本発明の第１の局面において、図１の装置は参照理由同定装置に対応し、被参照文書抽出手段１１、参照位置抽出手段１２、判断手段１３、および出力手段１４を備える。

【００１３】

被参照文書抽出手段１１は、与えられた文書データから被参照文書に関する文書情報を抽出する。参照位置抽出手段１２は、文書データ内で被参照文書を参照している位置に関する情報を抽出する。判断手段１３は、参照位置抽出手段１２により抽出された情報を解析して、被参照文書が参照されている理由を判断する。そして、出力手段１４は、被参照文書抽出手段１１により抽出された情報と被

参照文書が参照されている理由とを含む出力情報を出力する。

【００１４】

例えば、被参照文書は、与えられた文書データ内で参照されている参考文献等に対応し、被参照文書抽出手段１１は、その文書データから被参照文書のタイトル、著者名等の文書情報を抽出する。また、参照位置抽出手段１２は、文書データ内で参考文献の番号等を用いて被参照文書を参照している位置（参照位置）を探し、その位置が属する章、その位置の周辺の文字列等の情報を抽出する。

【００１５】

また、判断手段１３は、抽出された参照位置に関する情報に基づいて、被参照文書がその位置で参照されている理由（参照理由）を判断する。そして、出力手段１４は、被参照文書の文書情報と参照理由とを対応付けて出力する。

【００１６】

このような参照理由同定装置によれば、文書データから被参照文書の文書情報とその参照理由とが自動的に抽出され、利用者に提示される。したがって、利用者は、各被参照文書がどのような理由で参照元の文書により参照されているのかを認識することができ、参照理由から被参照文書の重要性を推定することもできる。さらに、利用者は、参照理由を利用して文書検索を行うこともできる。

【００１７】

また、本発明の第２の局面において、図１の装置は情報検索装置に対応し、文書データベース手段１５、参照関係格納手段１６、検索手段１７、および出力手段１４を備える。

【００１８】

文書データベース手段１５は、文書データを格納する。参照関係格納手段１６は、ある文書が被参照文書を参照している理由を表す参照理由を含む参照関係を格納する。検索手段１７は、参照関係格納手段１６に格納された参照関係を利用して、文書データベース手段１５に格納された文書データの検索を行う。そして、出力手段１４は、参照理由を含む検索結果を出力する。

【００１９】

文書データベース手段１５は、例えば、全文データベースに対応し、検索対象

となる文書データを格納する。参照関係格納手段１６は、単にある文書が他の文書を参照しているという参照関係ではなく、その参照理由をも含む参照関係の情報を格納する。検索手段１７は、格納された参照関係に含まれる参照理由を利用して、文書データベース手段１５の文書データを検索する。そして、出力手段１４は、検索に利用された参照理由に基づいて、得られた文書データの情報を出力する。

【００２０】

このような情報検索装置によれば、検索された被参照文書が参照理由とともに利用者に提示されるため、利用者は、参照理由から被参照文書の重要性を推定することができる。さらに、利用者は、希望条件に合った適当な参照理由を指定して文書検索を行うこともでき、検索が効率化される。

【００２１】

また、本発明の第３の局面において、図１の装置は文書分類装置に対応し、被参照文書抽出手段１１、参照位置抽出手段１２、判断手段１３、類似度判定手段１８、および出力手段１４を備える。

【００２２】

被参照文書抽出手段１１は、与えられた文書データから被参照文書に関する文書情報を抽出する。参照位置抽出手段１２は、文書データ内で被参照文書を参照している位置に関する情報を抽出する。判断手段１３は、参照位置抽出手段１２により抽出された情報を解析して、被参照文書が参照されている理由を判断する。類似度判定手段１８は、複数の文書データ間において、被参照文書が参照されている理由を含む参照関係の類似度を計算し、それらの文書データを分類する。そして、出力手段１４は、分類結果を出力する。

【００２３】

類似度判定手段１８は、例えば、２つの文書データ内で参照されている被参照文書の参照理由が与えられたとき、それらの参照理由を比較することで参照関係の類似度を計算する。このようにして、複数の文書データ内のすべての文書データ対について類似度が計算されると、それらの類似度に基づいて文書データを分類する。そして、出力手段１４は、文書クラスタ等の情報を分類結果として出力

する。

【００２４】

このような文書分類装置によれば、複数の文書データから被参照文書の文書情報とその参照理由とが自動的に抽出され、参照関係に基づいて複数の文書データが自動的に分類される。したがって、利用者は、参照理由を含む参照関係の観点から分類された文書群の情報を得ることができ、その情報を利用して文書検索を行うこともできる。

【００２５】

例えば、図１の被参照文書抽出手段１１は、後述する図２の文書構造解析部２２と参考文献解析部２５に対応し、図１の参照位置抽出手段１２は、図２の文書構造解析部２２と本文構造解析部２４に対応し、図１の判断手段１３は、図２の参照文脈解析部２６に対応し、図１の出力手段１４は、図２の対応判断部２７と出力成形部２８と出力インタフェース２９に対応する。

【００２６】

また、例えば、図１の文書データベース手段１５は、後述する図１１の全文データベース２に対応し、図１の参照関係格納手段１６は、図１１の参照関係データベース９５に対応し、図１の検索手段１７は、図１１の検索装置９２に対応する。

【００２７】

また、例えば、図１の類似度判定手段１８は、後述する図２３の類似度判定装置１５２に対応する。

【００２８】

【発明の実施の形態】

以下、図面を参照しながら、本発明の実施の形態を詳細に説明する。

一般に、システムが参照関係の特徴データを参照理由データ（関連性データ）としてあらかじめ持っていない場合、参照の有無を利用者に提示することは可能であるが、複数の参照理由を提示することは不可能か、あるいは非常に困難である。これに対して、システムが複数の参照理由データをあらかじめ持っている場合、参照の有無を利用者に提示することは簡単であり、複数の参照理由を提示す

ることも比較的簡単である。したがって、あらかじめ参照理由データをシステムに保持しておけば、それを利用して必要な情報を効率良く発見することができる。

【００２９】

まず、参照理由の同定に関しては、参照している文書内における被参照文書の出現位置、引用方法、文章等を解析することで、被参照文書が参照されている理由を判断する。この解析においては、例えば、以下のような特徴付けが行われる。

（１）被参照文献は、参照している文献の知識を補うものである。例えば、被参照文献が条例であり、それを参照している文献が条例の追加である場合が、これに相当する。また、被参照文献が星の発見に関する文献であり、それを参照している文献がその星に関する新たなデータや仮説等に関する文献である場合も、これに相当する。

（２）被参照文献は、参照している文献の扱う分野をまとめたものである。このような被参照文献は、いわゆるレビュー論文のようなものに相当し、その分野の近年の進歩状況を照会するためによく引用される）

（３）参照している文献が被参照文献に対する反論を行っている。

（４）被参照文献は、参照している文献で紹介している人物の代表的な文献である。

【００３０】

このようにして、数多くの文書を処理することにより、参照されている文書がその分野でどのような位置付けにあるのかを知ることができる。

図２は、参照理由同定装置の構成図である。図２の参照理由同定装置は、入力インタフェース２１、文書構造解析部２２、書誌情報解析部２３、本文構造解析部２４、参考文献情報解析部２５、参照文脈解析部２６、対応判断部２７、出力成形部２８、および出力インタフェース２９を備える。

【００３１】

まず、入力インタフェース２１は、文書データ３０をテキストデータとして入力する。文書構造解析部２２は、入力されたテキストデータを書誌情報、本文、

および参考文献の３つの部分に切り分けて、それぞれ書誌情報解析部２３、本文構造解析部２４、および参考文献解析部２５に渡す。

【００３２】

次に、書誌情報解析部２３は、書誌情報からタイトル、著者名等の文献情報を抽出して出力し、本文構造解析部２４は、本文から章構造と参考文献を引用している部分を抽出して出力し、参考文献解析部２５は、参考文献の記載から人名、発行年、タイトル、誌名等の文献情報を抽出して出力する。

【００３３】

また、参照文脈解析部２６は、本文中の参考文献を引用している部分を解析して参照理由を判断して出力し、対応判断部２７は、本文中の参考文献を引用している部分と参考文献の記載から抽出された文献情報とを対応付けて出力する。

【００３４】

次に、出力成形部２８は、書誌情報解析部２３から出力された文献情報と、参照文脈解析部２６から出力された参照理由と、対応判断部２７から出力された対応関係と、参考文献解析部２５から出力された参考文献の文献情報とをまとめて出力データ３１を成形し、出力インタフェース２９に渡す。

【００３５】

そして、出力インタフェース２９は、出力データ３１をディスプレイ画面等に出力する。出力データ３１には、例えば、文書データ３０の文献名と、参考文献および参照関係の組み合わせのリストが含まれる。

【００３６】

次に、図３から図９までを参照しながら、図２の参照理由同定装置の処理についてより詳細に説明する。

図３は、文書構造解析部２２の処理のフローチャートである。パターンデータリスト４１、４２、４３、４４は、あらかじめ決められた文字列パターンの情報を含んでおり、不図示の記憶装置上に保持される。

【００３７】

文書構造解析部２２は、まず、文書のテキストデータを読み込み（ステップＳ１）、パターンデータリスト４１に格納された導入部パターン（はじめにパター

ン）を参照しながら、マッチするパターンを含む行を探し（ステップＳ２）、そのような行がテキストデータにあるか否かをチェックする（ステップＳ３）。文書が英語または日本語で作成されている場合、導入部パターンとしては、例えば、次のような文字列がパターンデータリスト４１に格納される。

【００３８】

1 Introduction

1.Introduction

Introduction

１　はじめに

１．はじめに

はじめに

１　背景

概要

このようなパターンにマッチする行があれば、マッチする最初の行の位置をＰ１として記憶する（ステップＳ４）。そして、パターンデータリスト４２に格納されたキーワードパターンを参照しながら、その位置Ｐ１より前にキーワードパターンがあるか否かをチェックする（ステップＳ５）。キーワードパターンとしては、例えば、“＊＊＊”を任意のキーワードリストとして、次のような文字列がパターンデータリスト４２に格納される。

【００３９】

keyword ＊＊＊

keyword:＊＊＊

キーワード　＊＊＊

キーワード：＊＊＊

このようなパターンにマッチする行があれば、その位置を記憶する（ステップＳ６）。そして、パターンデータリスト４３に格納されたアブストラクトパターンを参照しながら、記憶された位置より前にアブストラクトパターンがあるか否かをチェックする（ステップＳ７）。アブストラクトパターンとしては、例えば、次のようなパターン情報がパターンデータリスト４３に格納される。

【００４０】

Abstract

２０語以上の文章で、文書の先頭の文章ではないもの。

このようなパターンにマッチする行があれば、その位置を記憶する（ステップＳ８）。そして、パターンデータリスト４４に格納された参考文献パターンを参照しながら、マッチするパターンを含む行を探し（ステップＳ９）、そのような行が位置Ｐ１より後の部分にあるか否かをチェックする（ステップＳ１０）。参考文献パターンとしては、例えば、次のような文字列がパターンデータリスト４４に格納される。

【００４１】

Reference

参照文献

参考文献

このようなパターンにマッチする行があれば、マッチする最後の行の位置をＰ２として記憶し（ステップＳ１１）、記憶した位置Ｐ１、Ｐ２等を出力して（ステップＳ１２）、処理を終了する。ステップＳ５においてキーワードパターンがない場合は、そのままステップＳ７以降の処理を行い、ステップＳ７においてアブストラクトパターンがない場合は、そのままステップＳ９以降の処理を行う。

【００４２】

また、ステップＳ３において導入部パターンがない場合は、文書の先頭の位置をＰ１として記憶し（ステップＳ１３）、ステップＳ９以降の処理を行う。また、ステップＳ１０において参考文献パターンがない場合は、文書の末尾の位置をＰ２として記憶し（ステップＳ１４）、ステップＳ１２の処理を行う。

【００４３】

ステップＳ１２において出力される位置Ｐ１は書誌情報の記載と本文の記載の境界位置（切れ目）に対応し、位置Ｐ２は本文の記載と参考文献の記載の境界位置に対応する。

【００４４】

次に、図４は、書誌情報解析部２３の処理のフローチャートである。ストップ

ワードリスト５１とパターンデータリスト５２は、あらかじめ決められた文字列パターンの情報を含んでおり、記憶装置上に保持される。

【００４５】

書誌情報解析部２３は、まず、テキストデータの先頭からＰ１までの部分を書誌情報として読み込み（ステップＳ２１）、ストップワードリスト５１を参照しながらテキストデータをチェックする。そして、最初にストップワードを含む文があればそれを読み飛ばし（ステップＳ２２）、次の文をタイトルとして抽出する（ステップＳ２３）。ストップワードとしては、タイトルではない文の先頭に現れる可能性のある文字列が用いられ、例えば、次のような文字列がストップワードリスト５１に格納される。

【００４６】

解説

技術メモ

Technical Note

次に、パターンデータリスト５２に格納された著者名パターンを参照しながら、タイトルとして記憶された文の次の文章からマッチするパターンを探し、それを抽出する（ステップＳ２４）。著者名パターンとしては、例えば、図４に示すようなパターン情報がパターンデータリスト５２に格納される。図４において、“○○”は辞書に登録された人名を表し、“－－－－－－－”は辞書に登録されていない未知語を表す。これらの著者名パターンは、次のように表すこともできる。

【００４７】

人名＋未知語

未知語＋未知語

アルファベット．人名

アルファベット．未知語

アルファベット．アルファベット．人名

アルファベット．アルファベット．未知語

次に、抽出されたデータをタイトルおよび著者名として記憶装置に書き出し（

ステップＳ２５）、処理を終了する。

【００４８】

次に、図５は、本文構造解析部２４の処理のフローチャートである。パターンデータリスト６１、６２は、あらかじめ決められた文字列パターンの情報を含んでおり、記憶装置上に保持される。

【００４９】

本文構造解析部２４は、文書構造解析部２２から渡された本文部分について、先頭から順に章構造を表すパターンを探していき、そのようなパターンの位置を記憶する処理を、それがなくなるまで繰り返す。次に、本文の先頭に戻って、参考文献の参照を表すパターンを探していき、そのようなパターンの位置を記憶する処理を、それがなくなるまで繰り返す。最後に、得られた章構造および参考文献の参照位置を出力する。

【００５０】

本文構造解析部２４は、まず、テキストデータのＰ１からＰ２までの部分を本文として読み込み（ステップＳ３１）、パターンデータリスト６１を参照しながら章構造を表すパターンを探していく（ステップＳ３２）。章構造のパターンとしては、例えば、次のようなパターン情報がパターンデータリスト６１に格納され、これらのパターンにマッチする行が探索される。

【００５１】

数字　文字列（改行）

数字．文字列（改行）

ここで、パターンの先頭の数字は章番号を表しており、通常、本文の後にいくに従って大きくなる。また、文字の大きさや太さ等があらかじめ分かっていれば、それらの情報も参考にして章構造のパターンが探索される。

【００５２】

次に、見つかったパターンに含まれる数字の増え方が適切か否かをチェックする（ステップＳ３３）。そして、例えば、その数字が前の数字に繋がっていなければ、数字の増え方が適切でないと判断し、エラー処理を行って（ステップＳ３４）、処理を終了する。

【００５３】

数字の増え方が適切であれば、見つかったパターンの位置を新たな章の先頭位置として記憶し（ステップＳ３５）、まだ文章が残っているか否かをチェックする（ステップＳ３６）。そして、文章が残っていれば、それがなくなるまでステップＳ３２以降の処理を繰り返す。

【００５４】

文章がなくなると、次に、位置Ｐ１にポインタをセットし（ステップＳ３７）、パターンデータリスト６２を参照しながら、参考文献を参照していることを表すパターンを探す（ステップＳ３８）。参考文献の参照を表すパターンとしては、例えば、次のようなパターン情報がパターンデータリスト６２に格納される。

【００５５】

数字）

〔数字〕

〔文字列　年号〕

文字列　年号

これらのパターンにマッチする部分が見つかれば、その位置を参照位置として記憶し、文章がまだ残っているか否かをチェックする（ステップＳ３９）。文章が残っていれば、それがなくなるまでステップＳ３８の処理を繰り返す。そして、文章がなくなると、得られた各章の先頭位置と参照位置を出力して（ステップＳ４０）、処理を終了する。

【００５６】

次に、図６は、参考文献解析部２５の処理のフローチャートである。パターンデータリスト７１は、あらかじめ決められた文字列パターンの情報を含んでおり、記憶装置上に保持される。

【００５７】

参考文献解析部２５は、文書構造解析部２２から渡された参考文献部分を１文ずつ読み込み、参考文献パターンとマッチングして、マッチした人名、発行年、タイトル、誌名等の情報を順に記憶していく。このような処理を参考文献部分の行がなくなるまで繰り返す。

【００５８】

参考文献解析部２５は、まず、テキストデータのＰ２から末尾までの部分を参考文献部分として読み込み（ステップＳ４１）、文章が残っているか否かをチェックする（ステップＳ４２）。文章が残っていれば、１文を読み込み（ステップＳ４３）、その文とパターンデータリスト７１の参考文献パターンとのマッチングを行い（ステップＳ４４）、マッチするパターンがあるか否かをチェックする（ステップＳ４５）。参考文献パターンとしては、例えば、次のようなパターン情報がパターンデータリスト７１に格納される。

【００５９】

人名　and　人名　年号　“タイトル”　誌名

〔人名　年号〕　人名　年号　“タイトル”　誌名

〔参照番号〕　人名　年号　“タイトル”　誌名

読み込んだ文がこのようなパターンにマッチすれば、その文に含まれる人名、年号、タイトル、誌名等の情報を参考文献情報として記憶し（ステップＳ４６）、ステップＳ４２以降の処理を繰り返す。また、マッチするパターンがなければ、その文に含まれる情報を記憶せずに、ステップＳ４２以降の処理を繰り返す。そして、ステップＳ４２において文章がなくなると、得られた参考文献情報を出力して（ステップＳ４７）、処理を終了する。

【００６０】

次に、図７は、参照文脈解析部２６の処理のフローチャートである。参照文脈解析部２６は、パターンデータリスト７１は、あらかじめ決められた文字列パターンの情報を含んでおり、記憶装置上に保持される。

【００６１】

参照文脈解析部２６は、言語解析部８１と参照特徴－参照理由対応表８２を含み、本文構造解析部２４から渡された章構造および参照位置の情報を解析して、参照理由を判断する。この参照特徴－参照理由対応表８２は、あらかじめ人手または学習により作成される。参照文脈解析部２６は、各参照位置について、その位置を含む章の情報と、その位置の周辺の文字列を言語解析部８１により解析した結果と、参考文献の引用パターンの３つの情報に基づき、参照特徴－参照理由

対応表８２を用いて参照理由を判断する。

【００６２】

参照文脈解析部２６は、まず、本文と章構造を読み込み（ステップＳ５１）、すべての参照位置について処理が完了したか否かをチェックする（ステップＳ５２）。処理が完了していなければ、次の参照位置を読み込み（ステップＳ５３）、章構造を参照しながらその位置が属する章の番号とその位置の周辺の文字列を抽出する（ステップＳ５４）。

【００６３】

次に、言語解析部８１は、抽出された文字列の形態素解析、構文解析、意味解析等を行って、参照位置周辺の特徴を抽出する（ステップＳ５５）。例えば、形態素解析については、以下の文献〔１〕のｐｐ．１１７－１３７のアルゴリズムが用いられ、構文解析については、この文献のｐｐ．１４０－１９９のアルゴリズムが用いられる。また、意味解析については、この文献のｐｐ．２００－２３１のアルゴリズムが用いられる。

〔１〕長尾真（ながおまこと），「自然言語処理」，岩波書店，１９９６

次に、参照文脈解析部２６は、参照特徴－参照理由対応表８２を参照しながら、参照理由を判断する（ステップＳ５６）。参照理由としては、例えば、以下のようなものが考えられる。

（１）反論（answer）

参考文献の内容と異なる意見を記述するために、参考文献を引用する。

（２）応用（application ）

応用分野を紹介するために、参考文献を引用する。

（３）基礎（basic ）

基礎的な研究や先人の仕事を紹介するために、参考文献を引用する。

（４）対立する意見（contraposition）

特定の内容と対立する意見や対照的な意見を紹介するために、参考文献を引用する。

（５）人物の代表的な文献（human ）

特定の人物の代表的な文献を紹介するために、参考文献を引用する。

（６）関連のある仕事（related work)

特定の内容と関連のある仕事を紹介するために、参考文献を引用する。

（７）まとめ（review)

特定の分野のまとめを紹介するために、参考文献を引用する。

（８）使用ソフトウェア（software)

シミュレーション等に使用したソフトウェアを紹介するために、参考文献を引用する。

（９）使用技術（technique ）

実験、シミュレーション等に使用した技術を紹介するために、参考文献を引用する。

（１０）弱い関係（weak correlation)

特定の内容と弱い関係を持つような内容を紹介するために、参考文献を引用する。

（１１）類似（similar ）

特定の内容と類似した内容を紹介するために、参考文献を引用する。

【００６４】

参照特徴－参照理由対応表８２には、このような参照理由のカテゴリと参照特徴との対応関係が格納されている。この参照特徴は、対応する参照理由を表現する本文の記述方法の特徴を表し、参考文献の参照位置を含む章の番号、参照位置の周辺の文字列の文脈、参考文献の引用パターン等の情報を含む。上述の参照理由の場合は、“＊＊”を参考文献を示す文字列として、例えば、以下のような対応関係が参照特徴－参照理由対応表８２に格納される。

【００６５】

参照位置＝１章、かつ、言語解析結果＝否定表現　→　反論

文脈＝“．．．＊＊というのが通説であるが、しかしながら、本研究では．．．”　→　反論

文脈＝“This approach is used in **...”　→　応用

参照位置＝１章または２章　→　基礎

文脈＝“The first idea...due to **”または“In previous research...**”　→　基礎

文脈＝“Unlike previous...**, new...”　→　対立する意見

参照位置＝１章、かつ、文脈＝“．．．が提案された＊＊”または“．．．が提案されている＊＊”　→　関連のある仕事

文脈＝“...is reviewd in ** ”または“** reviewd”または“see ** for an overview”　→　まとめ

文脈＝“We use...similar to **”または“see ** for a similar...approach ”または“...is also implemented in **”　→　類似

参照文脈解析部２６は、ステップＳ５４で抽出された情報やステップＳ５５で得られた解析結果を参照特徴として参照特徴－参照理由対応表８２を検索し、対応する参照理由を取得し、それを記憶する（ステップＳ５７）。そして、ステップＳ５２以降の処理を繰り返し、ステップＳ５２においてすべての参照位置についての処理が完了すると、各参照位置毎に参照理由を出力して（ステップＳ５８）、処理を終了する。

【００６６】

このように、参照理由は、参照を行っている文書中での被参照文書に関する表現の特徴に基づいて決められる。この参照理由を判断する方法としては、上述のような参照特徴－参照理由対応表８２を保持する方法のほかに、統計的な方法を用いることもできる。このような方法は機械学習とも呼ばれる。この場合、参照特徴としては、次のようなものが用いられる。

（１）参考文献の出現回数

（２）参考文献の参照位置：文書の先頭からの距離、出現した章

（３）参照位置の周辺情報：共起している単語、よく現れるフレーズ

これらの参照特徴に対して、被参照文書がどのような理由で参照されているのかを人間があらかじめ正解として用意しておき、機械もしくは人手で抽出した参照特徴を参照理由に対応付ける。このとき、参照特徴と参照理由のセットを複数用意しておき、統計的な方法により参照特徴から参照理由への対応を得る。この

統計的な方法としては、例えば、次の文献〔２〕のｐｐ．５２５－６５２に記載された決定木、ニューラルネット、最近傍法、ベイズ推定等の方法が挙げられる。

〔２〕S. Russell and P. Norvig, 「エージェントアプローチ　人工知能」, 共立出版, １９９７

次に、具体的な文書を例に挙げて、参照理由同定装置の動作を説明する。図８は、“Fujino”の文献を参考文献として参照している“Jpn.J.Appl.Phys.”の文書を示している。

【００６７】

まず、この文書が入力インタフェース２１を通して入力されると、入力されたデータは、文書構造解析部２２により、書誌情報（●Jpn.J.Appl.Phys.～KEYWORD;...）と本文（1 Introduction～）と参考文献（Reference～K.Fujino,...）の３つの部分に分割される。

【００６８】

ここでは、図３の分割位置Ｐ１は１章の“1 Introduction”の行の直前に設定され、分割位置Ｐ２は“Reference ”の行の直前に設定されている。したがって、“1 Introduction”の前の行までが書誌情報に対応し、１章から３章までが本文に対応し、“Reference ”の行以降が参考文献の部分に対応する。

【００６９】

このうち、書誌情報は書誌情報解析部２３により解析され、この文書の著者がKoji Tsukamotoらであること、タイトルが“Morphology Evolution....”であること等が同定される。また、本文は本文構造解析部２４により解析され、章構造が認識されるとともに、本文中で参考文献を引用している部分として、“6,8,10-13)”、“8,11,13)”のようなパターンが認識される。

【００７０】

このような認識結果は参照文脈解析部２６に送られ、参考文献がどのような章のどのような文章の中で参照されているかに基づいて、参照理由が判断される。

例えば、文献６）が１章で４回参照され、３章で２回参照されているものとすると、この文献は相対的に非常に多く参照されていることになる。このことから、文献６）は重要な文献であり、図８の文書の基礎となる文献であることが分かる。

【００７１】

また、文献６），８），１０－１３）は“Some methods have been.....However”という表現で参照されていることから、この文書とは対立する意見を述べた文献であることが分かる。また、文献８），１１），１３）は“Similar results were also reported..... ”という表現で参照されていることから、この文書と類似した文献であることが分かる。

【００７２】

また、参考文献の部分は参考文献解析部２５により解析され、参考文献の文献情報が抽出される。例えば、文献６）の作者は“Fujino”であり、発行年は“１９９１年”であり、掲載誌は“J.Electrochem.Soc.”であること等が認定される。

【００７３】

次に、対応判断部２７により、文献６）の参照を表している表記“$^{6)}$”と文献６）の文献情報とが対応付けられる。そして、出力成形部２８により、書誌情報解析部２３、参照文脈解析部２６、対応判断部２７、および参考文献解析部２５から出力された情報がまとめられて、出力インタフェース２９から出力される。

【００７４】

このようにして同定された参照理由を用いれば、ある文書が他の文献を参照している理由を利用して文書を検索することにより、検索を効率化することができる。

【００７５】

例えば、参照理由を用いない場合の参照関係のデータ構造は、図９に示すようになる。ここで、各アルファベット文字は１つの文書を表し、矢印の元の文書は矢印の先の文書を参照している。これに対して、参照理由を用いた場合の参照関係のデータ構造は、図１０に示すようになる。図１０では、文書間の参照／被参

照の情報だけでなく、参照理由毎の参照／被参照の情報が示されている。

【００７６】

図２７に示した従来の情報検索システムにおいて、参照関係データベースに図１０のような参照理由を用いた参照／被参照関係を格納した場合、システム構成は図１１のようになる。

【００７７】

図１１の情報検索システムは、入力装置９１、検索装置９２、表示装置９３、全文データベース９４、および参照関係データベース９５を備え、検索装置９２は、制御部１０１、検索部１０２、および選択部１０３を含む。参照関係データベース９５には、図２の参照理由同定装置により抽出された情報が格納される。

【００７８】

入力装置９１によりある文書が指定されると、制御部１０１による制御に基づいて、検索部１０２は、全文データベース９４の文書を検索する。そして、選択部１０３は、参照関係データベース９５の参照関係を用いて、指定された文書と共引用関係にある文献や似たような参照を行っている文献を、関連文書として選択し、表示装置９３は、それらの関連文書を表示する。

【００７９】

図１２は、このような検索結果の表示例を示している。参照関係データベース９５に指定された文書の参照関係１１１が格納されており、文書Ｘ、Ｗ、Ｖ等が関連文書として選択されたとする。このとき、表示装置９３の画面には検索結果１１２が表示され、各文書に該当する参照理由の欄にマーク“○”が記される。ここでは、関連文書が参照されている理由として、“answer”、“application”、“basic ”等が提示されている。

【００８０】

図１３は、図１１のシステムによる検索結果の表示処理のフローチャートである。選択部１０３は、まず、参照関係データベース９５から指定された文書の参照関係データを読み込み（ステップＳ６１）、そのデータに記述された参照理由を読み込む（ステップＳ６２）。次に、指定された文書が参照している被参照文書の文献情報を読み込んで、図１２のような表形式の検索結果を表示し（ステッ

プＳ６３）、該当する参照理由の欄にマークを付ける（ステップＳ６４）。

【００８１】

次に、参照関係データが残っているか否かをチェックし、データが残っていれば、ステップＳ６１以降の処理を繰り返す。そして、指定された文書のすべての参照関係データが処理されると、表示処理を終了する。

【００８２】

このような表示処理によれば、ある文書に関連する重要文書を表示する際、図２８に示したような単一の基準（参照回数）に基づく順序だけでなく、その重要文書が参照されている理由も明示される。したがって、参照理由に基づいて検索結果をさらに絞り込むことも可能になり、文書検索が効率化される。

【００８３】

また、関連文書を単に重要度順に表示するだけでなく、図１４に示すように、関連文書を参照理由毎に分類して重要度順に表示することも可能である。この場合、各文書の重要度は、上記単一の基準に基づく重要度に加え、以下のような要素を加味して決定される。

（１）それぞれの参照理由

（２）対応する参照理由により参照された回数

（３）参照元の文書の重要度

（４）出典の重要度

（５）利用者に参照された回数

表示された参照理由の中から利用者が適当なものを選択して、さらに検索を指示した場合、検索部１０２は、参照関係データベース９５を参照しながら全文データベース９４を検索する。そして、指定された参照理由で参照されている文書が関連文書として表示される。

【００８４】

このように、図１１の情報検索システムによれば、検索結果に参照理由を付与したり、参照理由を用いて検索を行ったりすることで、利用者が情報検索の効率を改善することができる。

【００８５】

次に、被参照文書に関する情報を抽出して提示する情報提示装置について説明する。図１５は、このような情報提示装置の構成図である。図１５の情報提示装置は、図１１の情報検索システムの構成要素に加えて、入力インタフェース１２１、検索入力インタフェース１２２、検索出力インタフェース１２３、参照理由同定装置１２４、および参照関係変換装置１２５をさらに備える。

【００８６】

入力インタフェース１２１は、ある文書を参照している複数の文書（文書群）を入力し、全文データベース９４に格納するとともに、参照理由同定装置１２４に与える。参照理由同定装置１２４は、図２のような構成を持ち、与えられた文書の参照関係を求めて、参照関係変換装置１２５に出力する。

【００８７】

参照関係変換装置１２５は、与えられた参照関係を変換して、参照関係データベース９５に格納する。例えば、図１６に示すように、文書Ａが文書Ｂ、Ｃ、Ｄをそれぞれｂ、ｃ、ｄという参照理由で参照しているという参照関係が与えられたとする。このとき、参照関係変換装置１２５は、図１７に示すように、文書Ｂ、Ｃ、Ｄが文書Ａからそれぞれｂ、ｃ、ｄという理由で参照されていることを表すデータを加えて、図１８に示すような参照関係データを生成する。そして、このデータを参照関係データベース９５に格納する。

【００８８】

利用者は、検索入力インタフェース１２２から、キーワードや文書を指定して検索を行う。そして、検索出力インタフェース１２３により、検索結果の文書の情報が表示されると、その文書の参照理由を用いてさらに検索を行う。

【００８９】

検索入力インタフェース１２２および検索出力インタフェース１２３では、以下のような入出力が行われる。

（１）利用者が文書を指定することで、その文書の本文を表示する。

（２）興味のある参照関係をフィルタとして使用する。

（３）参照理由を明示的に表示して参照関係を表示する。

（４）ＧＵＩ（graphical user interface）を用いて、参照理由を明示しながら

、あるいは参照関係を表す線または矢印の種類を参照理由によって変えながら、参照関係を画像表示する。

【００９０】

例えば、複数の文書を発行年順に並べて、文書間の参照関係を矢印で表すと、図１９のようになる。ここでは、円形のマークが１つの文書を表し、複数の文書が時間軸に沿って時系列に表示されている。また、参照元の文書と被参照文書は参照理由毎に異なる線種の矢印で結ばれる。

【００９１】

このようなＧＵＩを用いれば、複雑な参照関係における各文書の位置付けを明確に認識することができ、文書検索が効率化される。さらに、この表示画面中の文書マークを利用者が指定することにより、対応する文書の内容を表示したり、その文書が参照している文書のリストを表示したりすることもできる。

【００９２】

図１９では、矢印の線種により参照理由を区別しているが、矢印の色によりこれを区別してもよく、参照理由を文字列で表示してもよい。

次に、参照関係を利用したキーワード抽出装置について説明する。図２０は、このようなキーワード抽出装置の構成図である。図２０のキーワード抽出装置は、図２の参照理由同定装置の構成要素に加えて、キーワード抽出部１３１をさらに備える。

【００９３】

本文構造解析部２４は、解析結果をキーワード抽出部１３１に出力し、キーワード抽出部１３１は、本文中で参考文献が参照されている位置の周辺の文章からキーワードを抽出し、それを参考文献のキーワードとして出力する。この場合、出力インタフェース２９から出力される出力データ１３２には、文書データ３０の文献名に加えて、参考文献、参照関係、およびキーワードの組み合わせのリストが含まれる。

【００９４】

このようなキーワード抽出装置によれば、文献の著者自身が選択したキーワードだけでなく、その文献を参照している他の文書に基づくキーワードも付与され

ることになる。したがって、主観的なキーワードだけではなく、客観的なキーワードを付与することが可能になる。

【００９５】

図２１は、このような参照関係を利用したキーワードの例を示している。ここでは、ある文献の著者が“Machine Learning”、“Decision Tree ”等をキーワードとして付与し、キーワード抽出装置は、その文献を参照している他の文書の文脈から“Corpus”という新たなキーワードを抽出してキーワードのデータ構造に付加している。

【００９６】

図２２は、このようにして付与されたキーワードを用いて文書を検索する情報検索システムを示している。図２２の情報検索システムは、図１１に示したシステムの構成要素に加えて、文献キーワード記憶部１４１をさらに備える。文献キーワード記憶部１４１には、文献にあらかじめ付与されたキーワードが格納され、参照関係データベース９５には、図２０のキーワード抽出装置により抽出された情報が格納される。

【００９７】

検索部１０２は、文献キーワード記憶部１４１と参照関係データベース９５のキーワードを参照しながら全文データベース９４の文書を検索し、選択部１０３は、文献キーワード記憶部１４１と参照関係データベース９５の情報を用いて関連文書を選択する。このとき、表示の優先順位は、キーワードの一致度、参照関係に基づく重要度、および文書に対するアクセス回数を加味して決められる。

【００９８】

このような情報検索システムによれば、あらかじめ付与されたキーワードに加えて、参照関係を用いて客観的に付与されたキーワードを利用して検索することができ、検索効率が向上する。

【００９９】

また、文書間の参照関係を利用して多数の文書を分類することも可能である。図２３は、このような参照関係に基づく分類を行う文書分類装置の構成図である。図２３の文書分類装置は、参照関係データベース９５、キーワード抽出装置１

５１、および類似度判定装置１５２を備える。

【０１００】

キーワード抽出装置１５１は、図２０のような構成を持ち、複数の文書データ１５３から参考文献に関する参照関係、キーワード等の情報を抽出して、参照関係データベース９５に格納する。類似度判定装置１５２は、入力された文書間で参照理由を含む参照関係を比較し、参照関係の類似度に基づくクラスタリングを行って、文書クラスタ１５４を出力する。

【０１０１】

例えば、文書ａと文書ｂの間で参照関係の類似度を計算する際に、次のような計算式が用いられる。

【０１０２】

【数１】

$$sim(a, b) = \frac{1}{\sqrt{n_a}\sqrt{n_b}} \sum_{i}^{n_a} \sum_{j}^{n_b} \delta(r_{ai}, r_{bj})$$

【０１０３】

この計算式において、ｓｉｍ（ａ，ｂ）は、文書ａと文書ｂの間の類似度を表す。また、ｎａ、ｎｂは、それぞれ、文書ａ、文書ｂが参照している被参照文書の数であり、ｒａｉ、ｒｂｊは、それぞれ、文書ａ、文書ｂによる参照を表すベクトルであり、被参照文書とその参照理由の属性を持つ（ｉ＝１，２，．．．，ｎａ，ｊ＝１，２，．．．，ｎｂ）。

【０１０４】

δ（ｒａｉ，ｒｂｊ）は、ｒａｉとｒｂｊの間の類似度を表す関数である。この関数は、例えば、ｒａｉとｒｂｊが同じ被参照文書を同じ参照理由で参照している場合はδ（ｒａｉ，ｒｂｊ）＝１となり、そうでない場合はδ（ｒａｉ，ｒｂｊ）＝０となるように定義される。また、ｒａｉとｒｂｊが同じ被参照文書を異なる参照理由で参照していた場合に、δ（ｒａｉ，ｒｂｊ）に１または０．５を割り当てるという方法もある。

【０１０５】

被参照文書についても同じ方法で類似度計算が行われ、計算結果に基づいて文書のクラスタリングが行われる。文書のクラスタリングには、例えば、上述の文献〔１〕のｐｐ．４３６－４３８に記載されているアルゴリズムが用いられる。これにより、参照関係が互いに類似している複数の文書が同じクラスタに分類される。また、このような類似度計算に、さらにキーワードの一致度、単語の出現頻度等を加味したクラスタリングを行うこともできる。

【０１０６】

このように、文書間の参照関係に基づいて文書を分類することにより、利用者は、参照理由を含む参照関係の観点から分類された文書群の情報を得ることができる。また、その分類結果を利用して文書を検索すれば、検索が効率化される。

【０１０７】

図２４は、図１９に示した参照関係の時系列表示において、文書の分類結果を利用した例を示している。図２４において、縦軸は、分類により得られた文書の種類を表し、横軸は、時間を表す。このような表示方法によれば、同じような参照関係を持つ文書が近接して配置され、多数の文書間の参照関係をより分かりやすく表示することができる。

【０１０８】

以上説明した実施形態においては、文書間の参照関係に関する処理を行っているが、同様にして、任意の情報間における参照関係を処理することもできる。例えば、参考文献の代わりに、他のテキストデータ、画像データ、音声データ、プログラムリスト等が参照されている場合、それらの情報の参照理由が同定され、情報検索に利用される。

【０１０９】

ところで、図２の参照理由同定装置、図１１および図２２の情報検索システム、図１５の情報提示装置、図２０のキーワード抽出装置、および図２３の文書分類装置は、図２５に示すような情報処理装置（コンピュータ）を用いて構成することができる。図２５の情報処理装置は、ＣＰＵ（中央処理装置）１６１、メモリ１６２、入力装置１６３、出力装置１６４、外部記憶装置１６５、媒体駆動装置１６６、およびネットワーク接続装置１６７を備え、それらはバス１６８によ

り互いに接続されている。

【０１１０】

メモリ１６２は、例えば、ＲＯＭ（read only memory）、ＲＡＭ（random access memory）等を含み、処理に用いられるプログラムとデータを格納する。ＣＰＵ１６１は、メモリ１６２を利用してプログラムを実行することにより、必要な処理を行う。

【０１１１】

入力装置１６３は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、利用者からの指示や情報の入力に用いられる。出力装置１６４は、例えば、ディスプレイ、プリンタ、スピーカ等であり、利用者へのメッセージや処理結果の出力に用いられる。

【０１１２】

外部記憶装置１６５は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク（magneto-optical disk）装置等であり、図１１の全文データベース９４、参照関係データベース９５、および図２２の文献キーワード記憶部１４１として用いられる。また、情報処理装置は、この外部記憶装置１６５に、上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ１６２にロードして使用することができる。

【０１１３】

媒体駆動装置１６６は、可搬記録媒体１６９を駆動し、その記録内容にアクセスする。可搬記録媒体１６９としては、メモリカード、フロッピーディスク、ＣＤ－ＲＯＭ（compact disk read only memory）、光ディスク、光磁気ディスク等、任意のコンピュータ読み取り可能な記録媒体が用いられる。利用者は、この可搬記録媒体１６９に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ１６２にロードして使用することができる。

【０１１４】

ネットワーク接続装置１６７は、任意のネットワーク（回線）を介して外部の装置と通信し、通信に伴うデータ変換を行う。情報処理装置は、必要に応じて、ネットワーク接続装置１６７を介して上述のプログラムとデータを外部の装置か

ら受け取り、それらをメモリ１６２にロードして使用することができる。

【０１１５】

図２６は、図２５の情報処理装置にプログラムとデータを供給することのできるコンピュータ読み取り可能な記録媒体を示している。可搬記録媒体１６９や外部のデータベース１７０に保存されたプログラムとデータは、メモリ１６２にロードされる。そして、ＣＰＵ１６１は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【０１１６】

【発明の効果】

本発明によれば、文書等の情報内で他の情報を参照している部分を解析することにより、その情報が参照されている理由を判断することができる。また、情報の参照理由を利用者に提示すれば、利用者は複数の参照理由から希望するものを選択することができ、情報検索が効率化される。さらに、参照理由を用いて情報間の参照関係を提示することにより、各情報の位置付けが明確に認識されるようになる。

【図面の簡単な説明】

【図１】

本発明の装置の原理図である。

【図２】

参照理由同定装置の構成図である。

【図３】

文書構造解析部の処理のフローチャートである。

【図４】

書誌情報解析部の処理のフローチャートである。

【図５】

本文構造解析部の処理のフローチャートである。

【図６】

参考文献解析部の処理のフローチャートである。

【図７】

参照文脈解析部の処理のフローチャートである。

【図８】

文書の例を示す図である。

【図９】

第１の参照関係を示す図である。

【図１０】

第２の参照関係を示す図である。

【図１１】

情報検索システムの構成図である。

【図１２】

第１の検索結果表示を示す図である。

【図１３】

表示処理のフローチャートである。

【図１４】

第２の検索結果表示を示す図である。

【図１５】

情報提示装置の構成図である。

【図１６】

第３の参照関係を示す図である。

【図１７】

第４の参照関係を示す図である。

【図１８】

第５の参照関係を示す図である。

【図１９】

第１の時系列表示を示す図である。

【図２０】

キーワード抽出装置の構成図である。

【図２１】

参照関係を用いたキーワードを示す図である。

【図２２】

キーワードを用いた検索を示す図である。

【図２３】

文書分類装置の構成図である。

【図２４】

第２の時系列表示を示す図である。

【図２５】

情報処理装置の構成図である。

【図２６】

記録媒体を示す図である。

【図２７】

従来の情報検索システムを示す図である。

【図２８】

従来の表示形式を示す図である。

【符号の説明】

１、９２　検索装置

２、９４　全文データベース

３、９５　参照関係データベース

４、１０１　制御部

５、１０２　検索部

６、１０３　選択部

１１　被参照文書抽出手段

１２　参照位置抽出手段

１３　判断手段

１４　出力手段

１５　文書データベース手段

１６　参照関係格納手段

１７　検索手段

１８　類似度判定手段

２１、１２１　入力インタフェース

２２　文書構造解析部

２３　書誌情報解析部

２４　本文構造解析部

２５　参考文献解析部

２６　参照文脈解析部

２７　対応判断部

２８　出力成形部

２９　出力インタフェース

３０、１５３　文書データ

３１、１３２　出力データ

４１、４２、４３、４４、５２、６１、６２、７１　パターンデータリスト

５１　ストップワードリスト

８１　言語解析部

８２　参照特徴－参照理由対応表

９１、１６３　入力装置

９３　表示装置

１１１　参照関係

１１２　検索結果

１２２　検索入力インタフェース

１２３　検索出力インタフェース

１２４　参照理由同定装置

１２５　参照関係変換装置

１３１　キーワード抽出部

１４１　文献キーワード記憶部

１５１　キーワード抽出装置

１５２　類似度判定装置

１５４　文書クラスタ

１６１　ＣＰＵ

１６２　メモリ

１６４　出力装置

１６５　外部記憶装置

１６６　媒体駆動装置

１６７　ネットワーク接続装置

１６８　バス

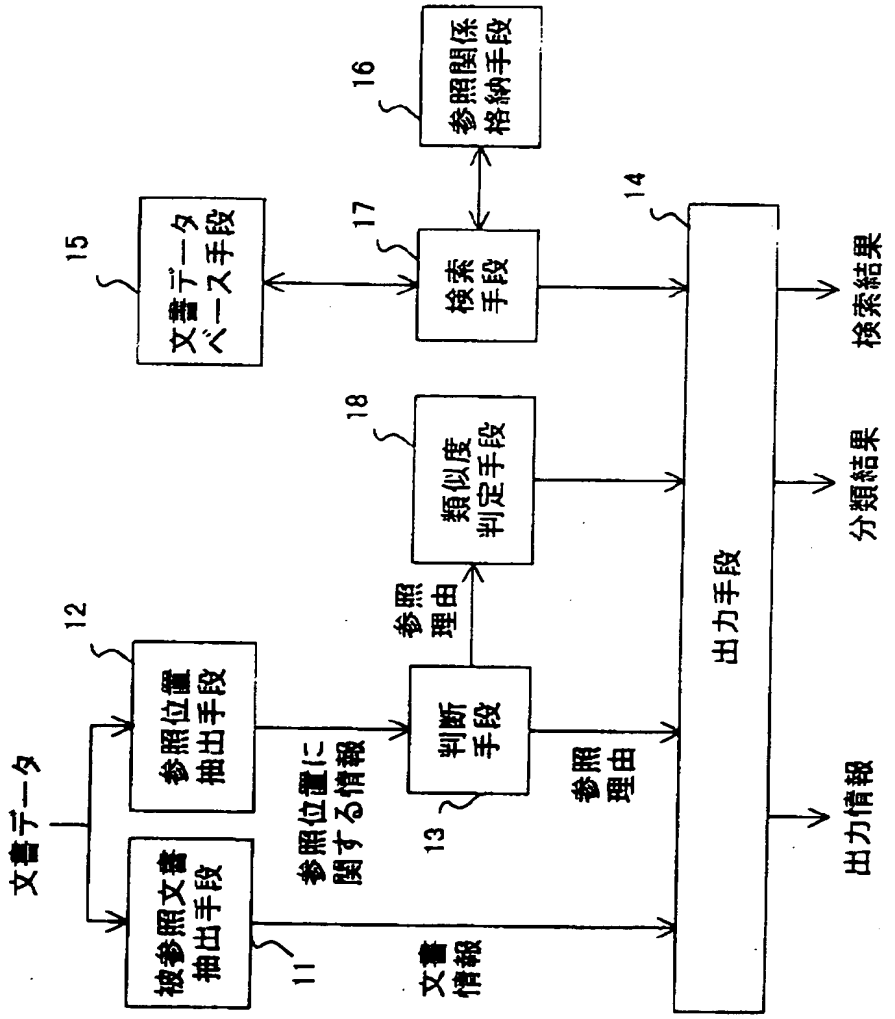
１６９　可搬記録媒体

１７０　データベース

【書類名】　　図面

【図１】

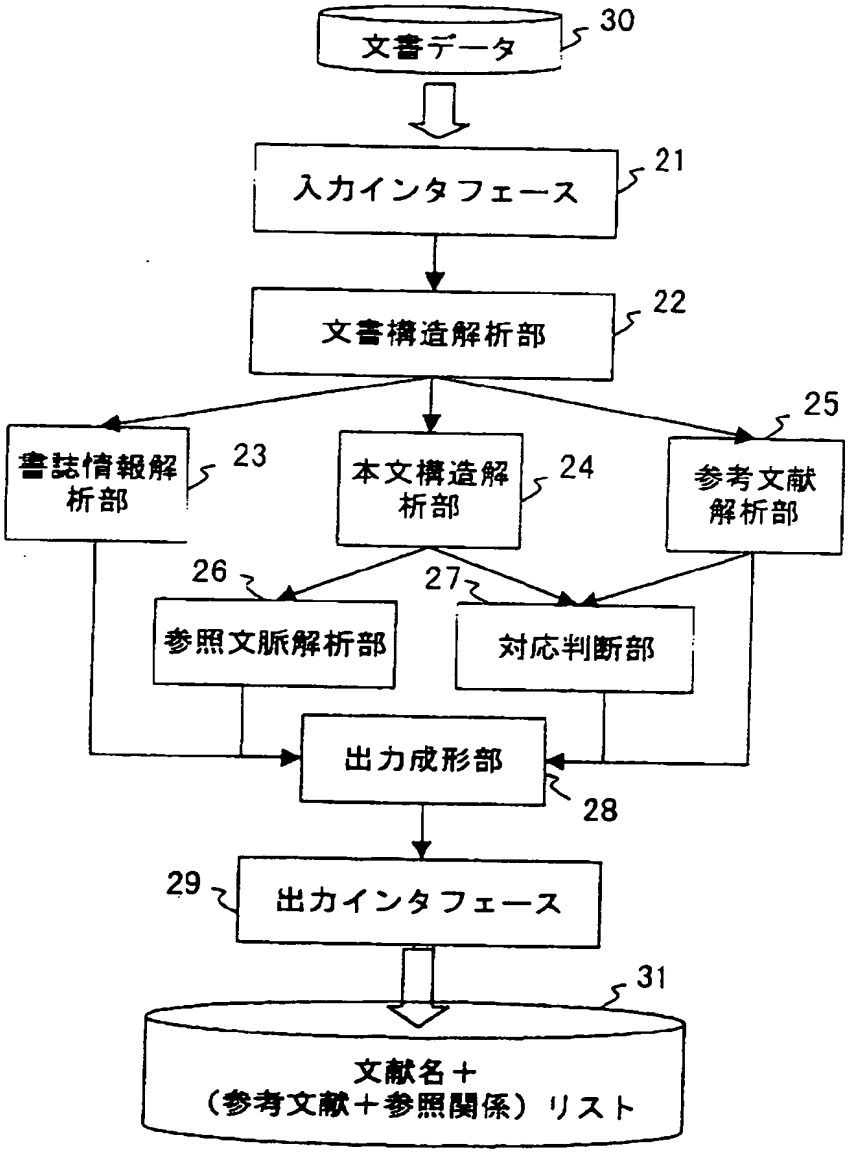
本　発　明　の　原　理　図



出証特２０００－３０３１６６４

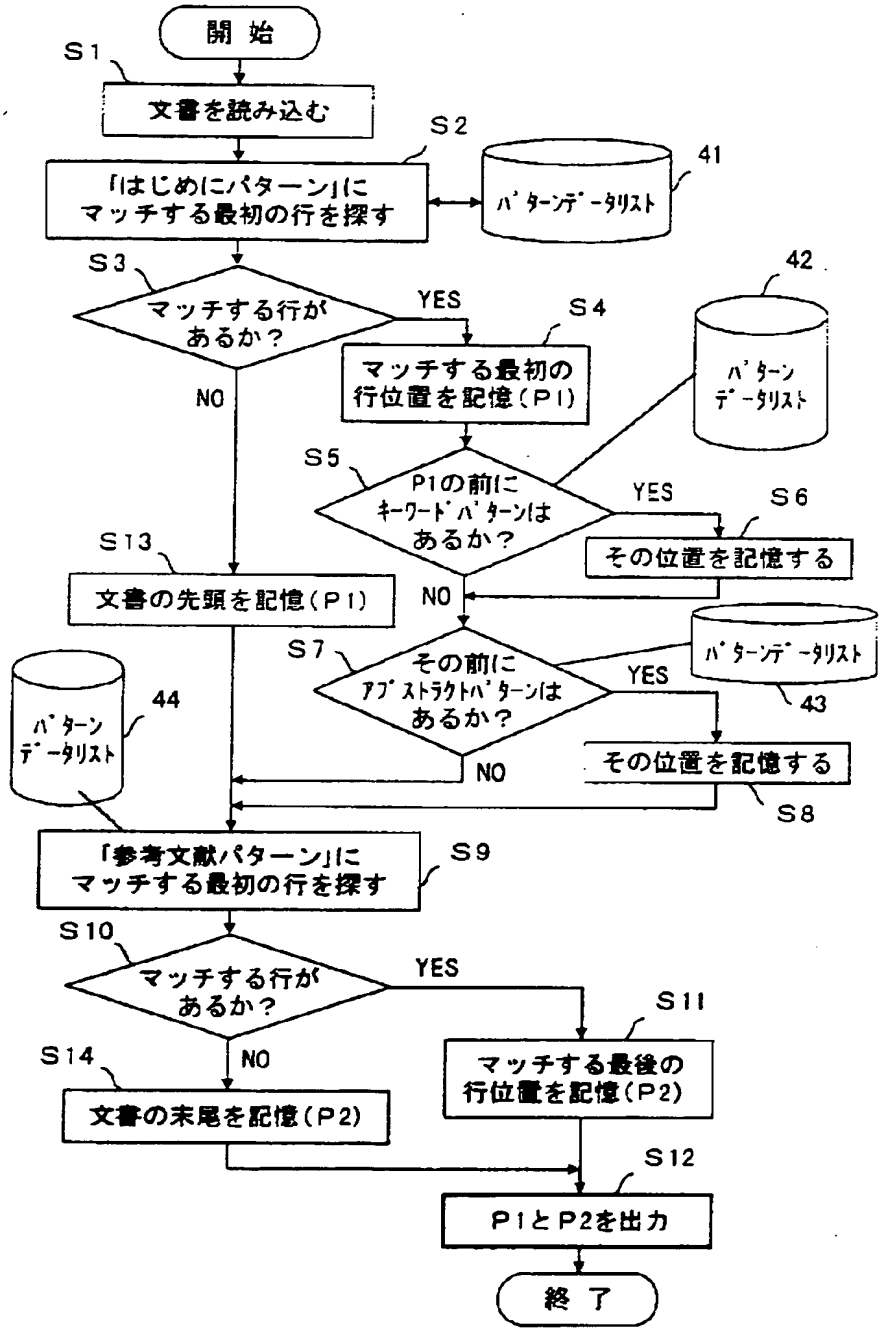
【図２】

参照理由同定装置の構成図



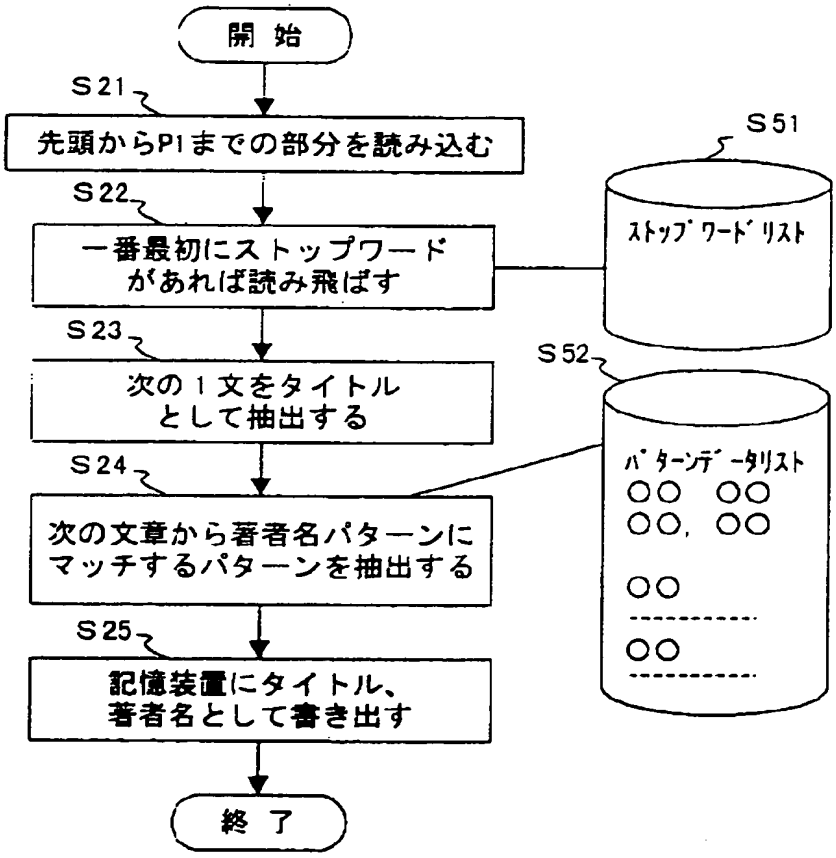
【図３】

文書構造解析部の処理のフローチャート



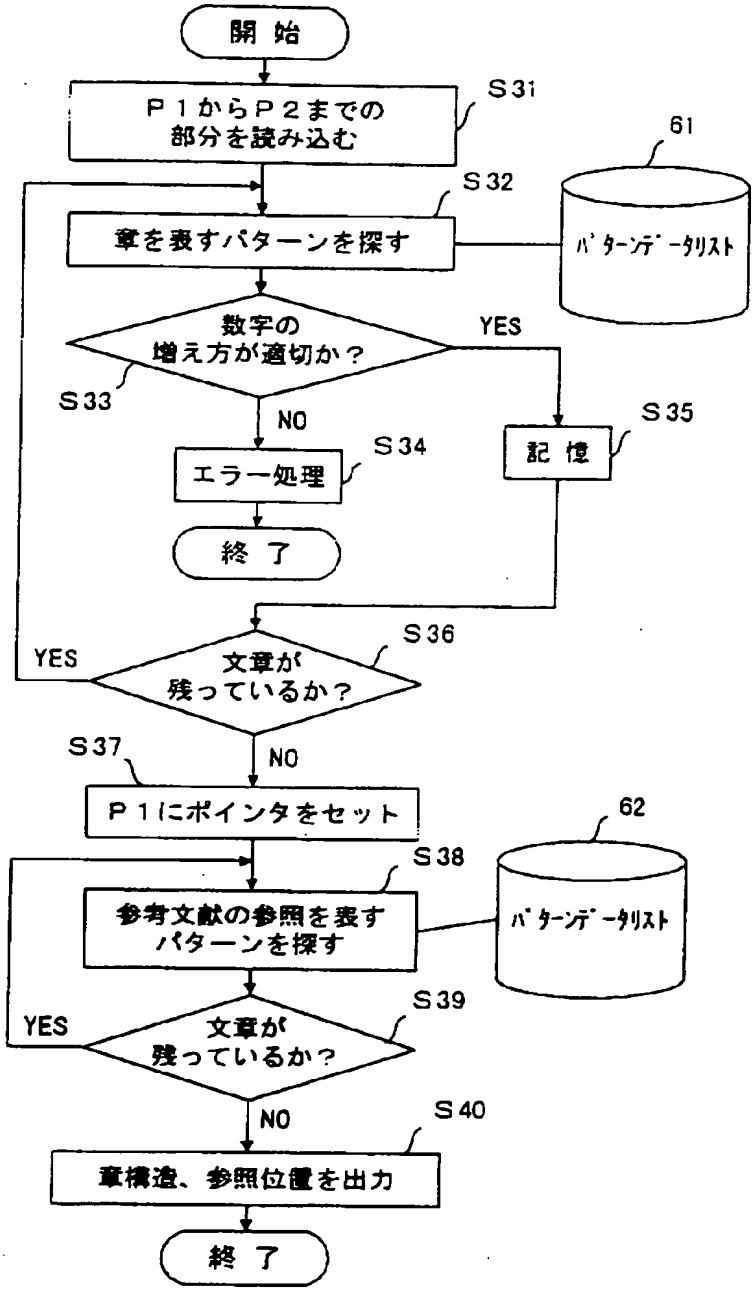
【図４】

書誌情報解析部の処理のフローチャート



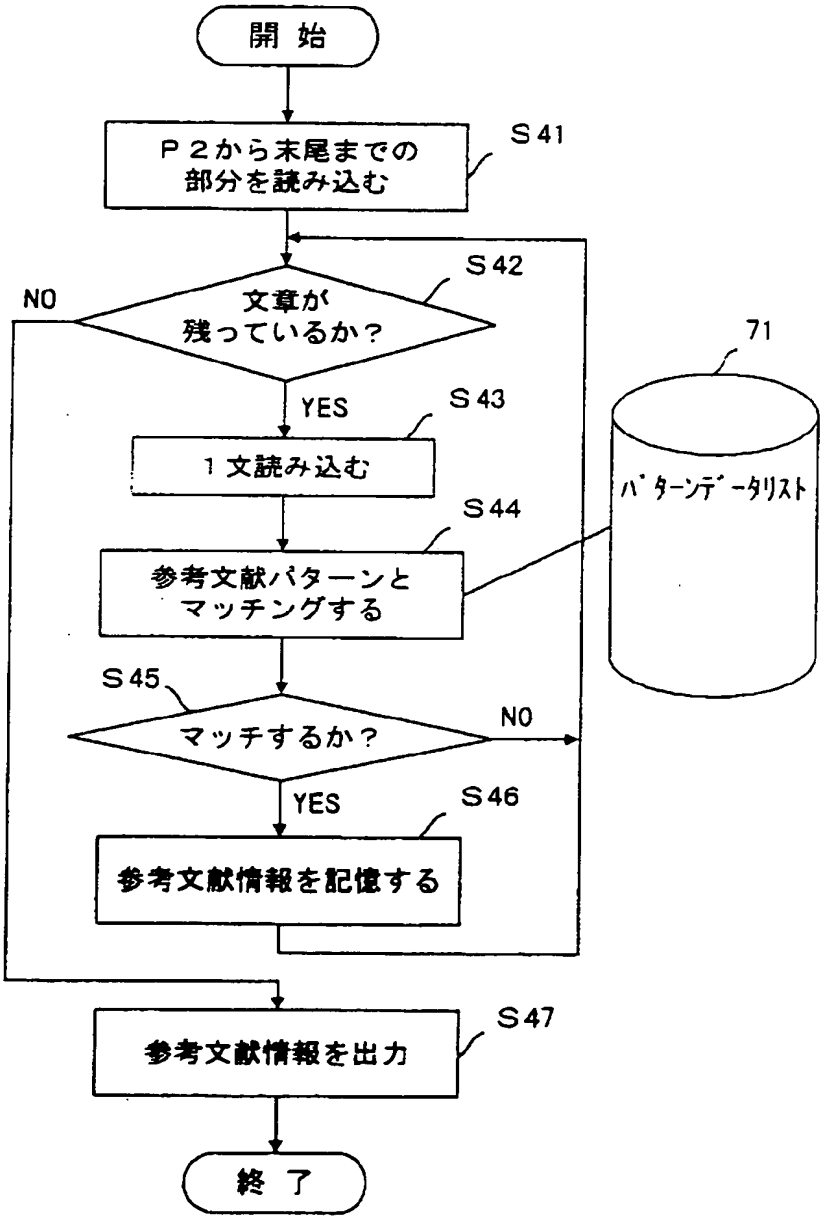
【図５】

本文構造解析部の処理のフローチャート



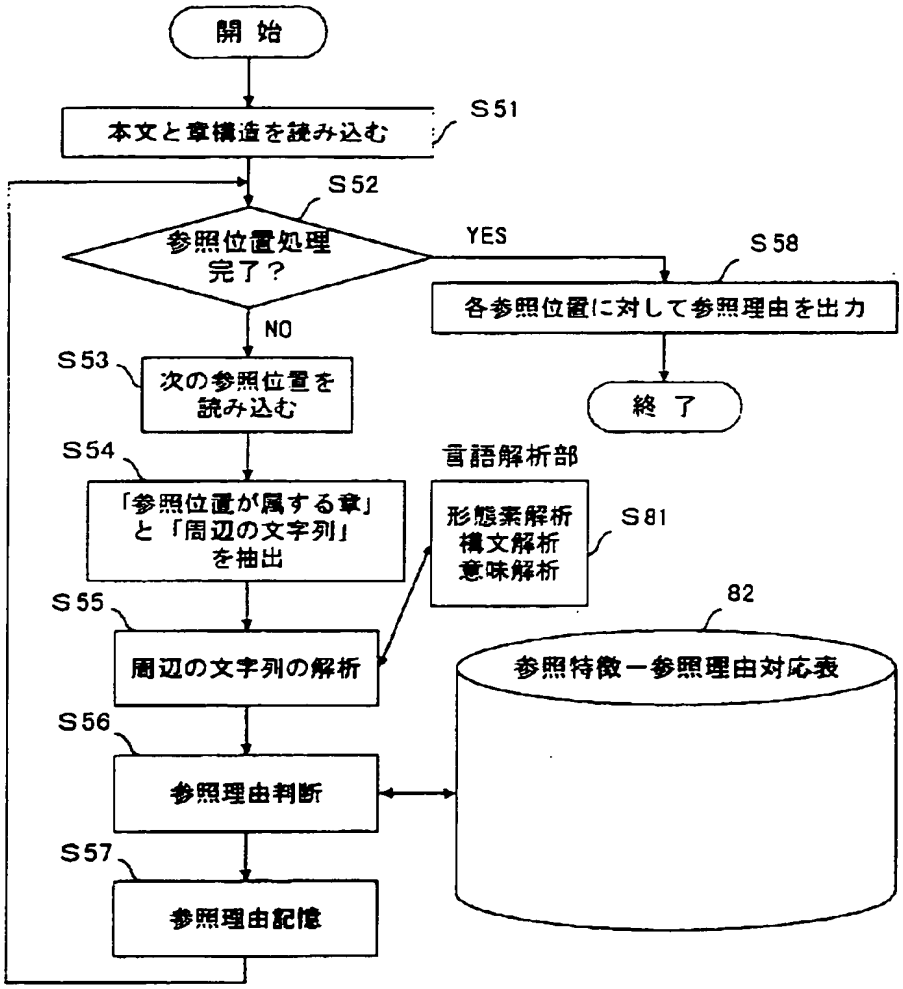
【図６】

参考文献解析部の処理のフローチャート



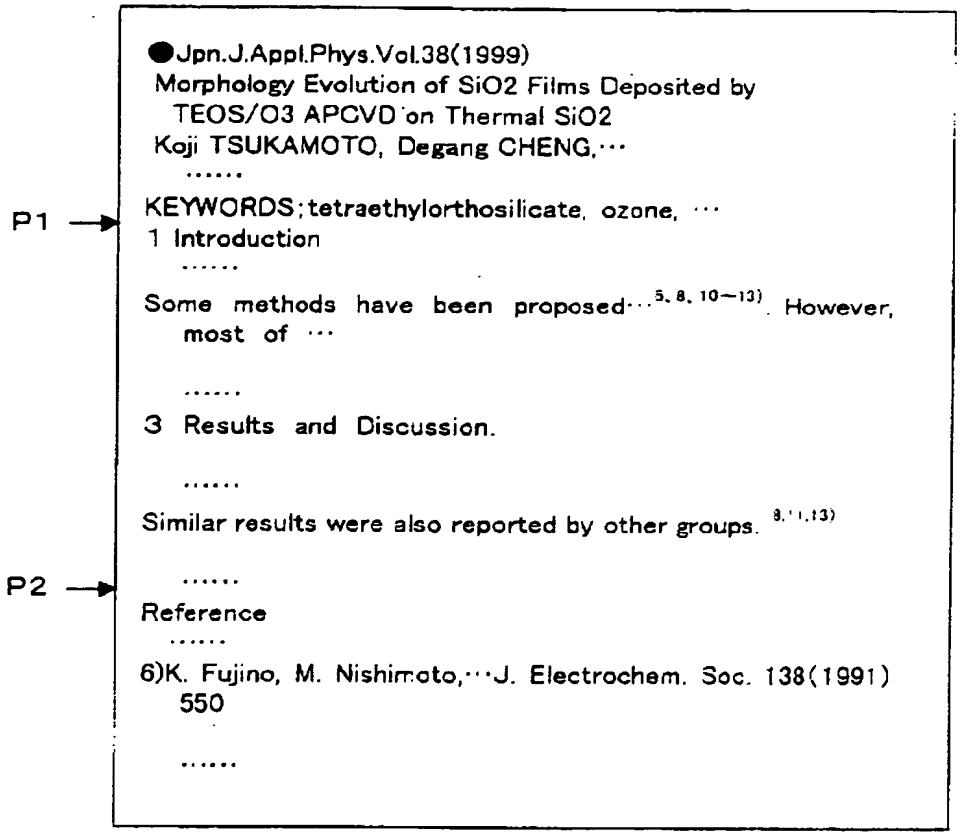
【図７】

参照文脈解析部の処理のフローチャート



【図８】

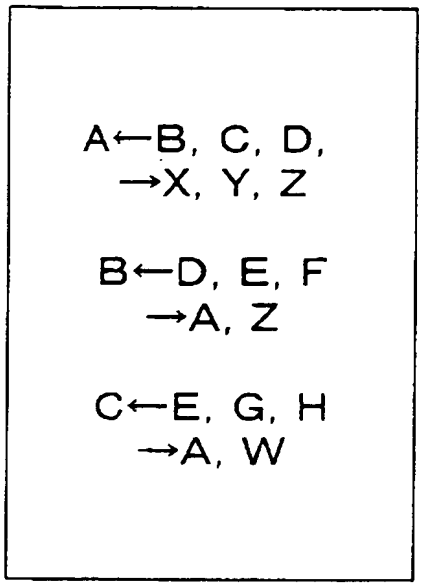
文書の例を示す図



(Koji Tsukamoto et. al. (1999), *Morphology Evolution of SiO2 Films Deposited by TEOS/O3 APCVD on Thermal SiO2*, Jpn. J. Appl. Phys.)

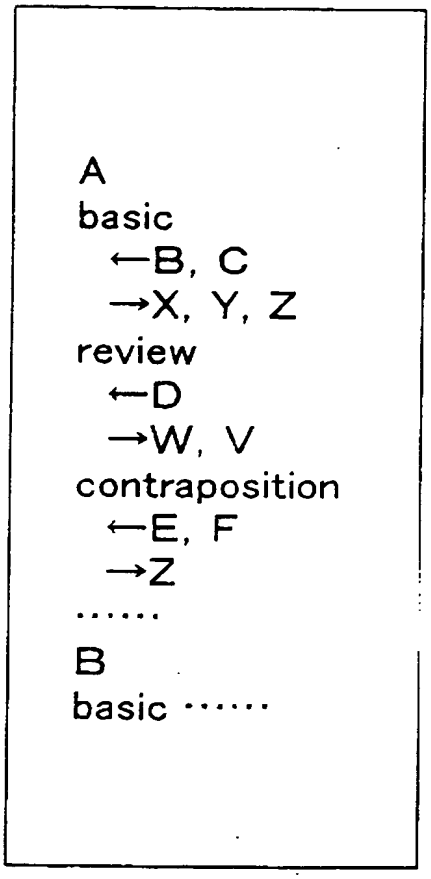
【図９】

第1の参照関係を示す図



【図１０】

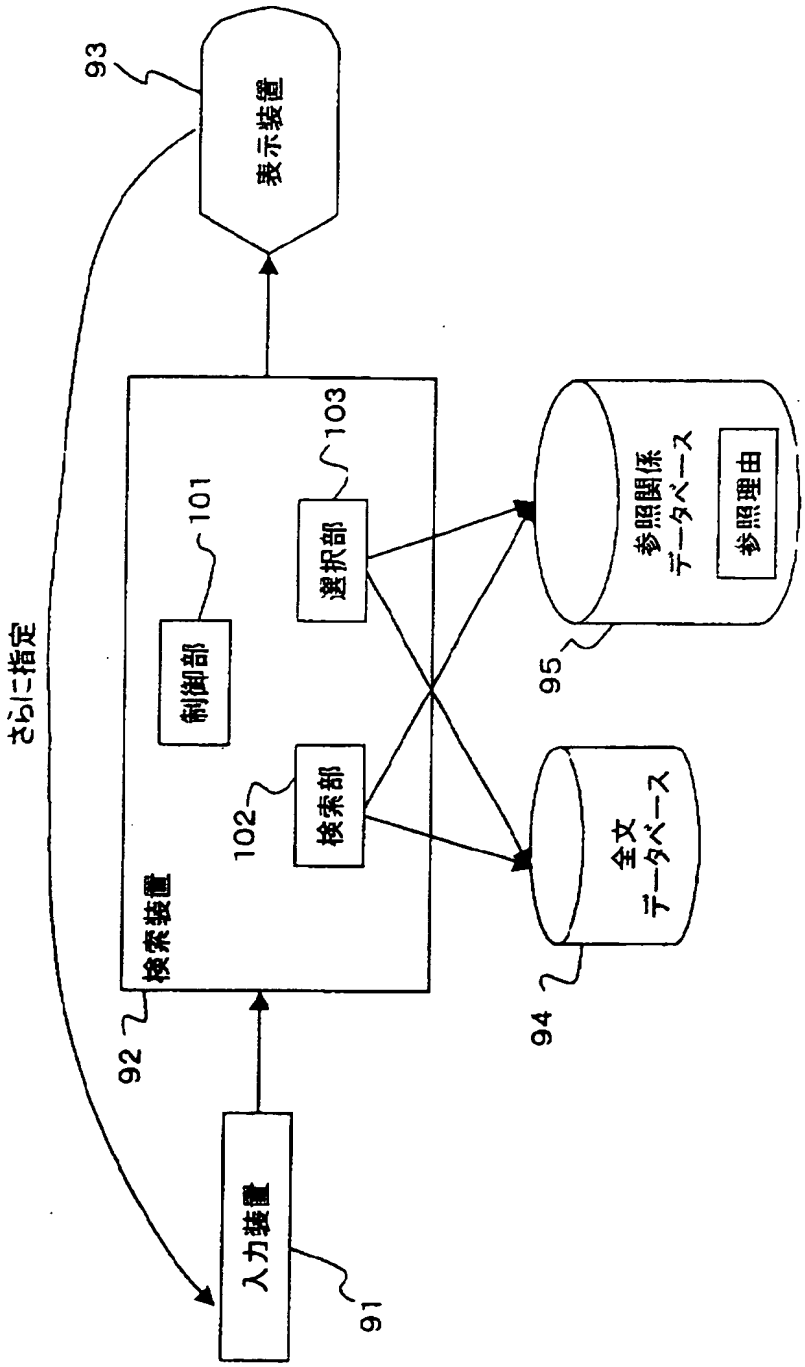
第2の参照関係を示す図



【図１１】

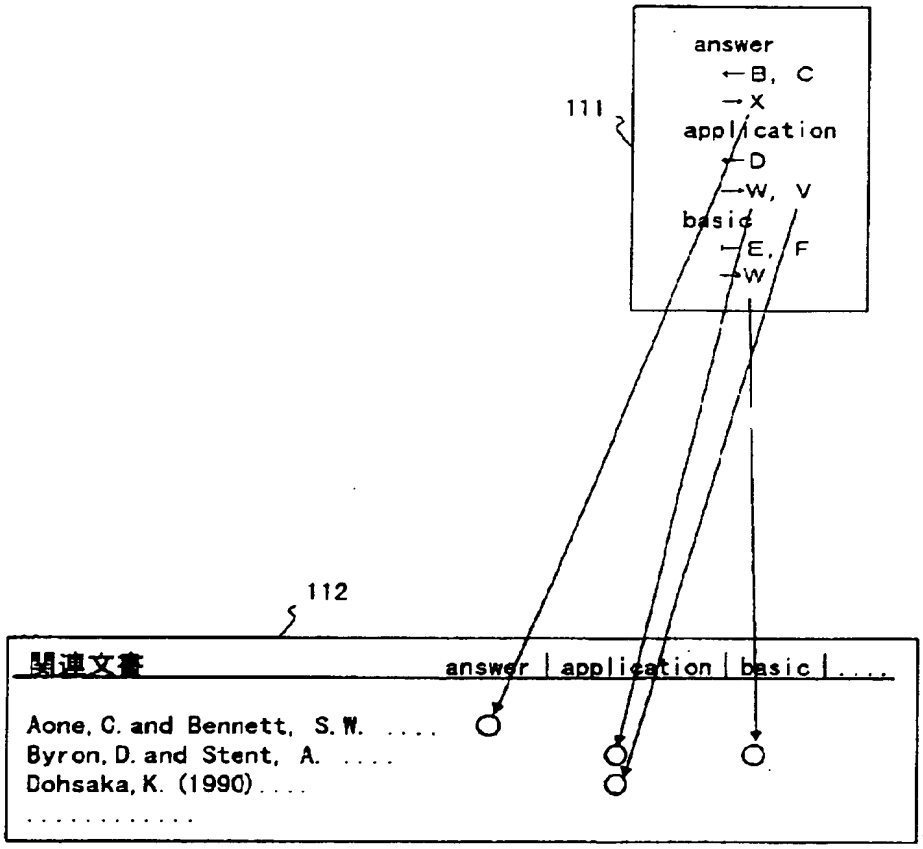
出証特２０００－３０３１６６４

情報検索システムの構成図



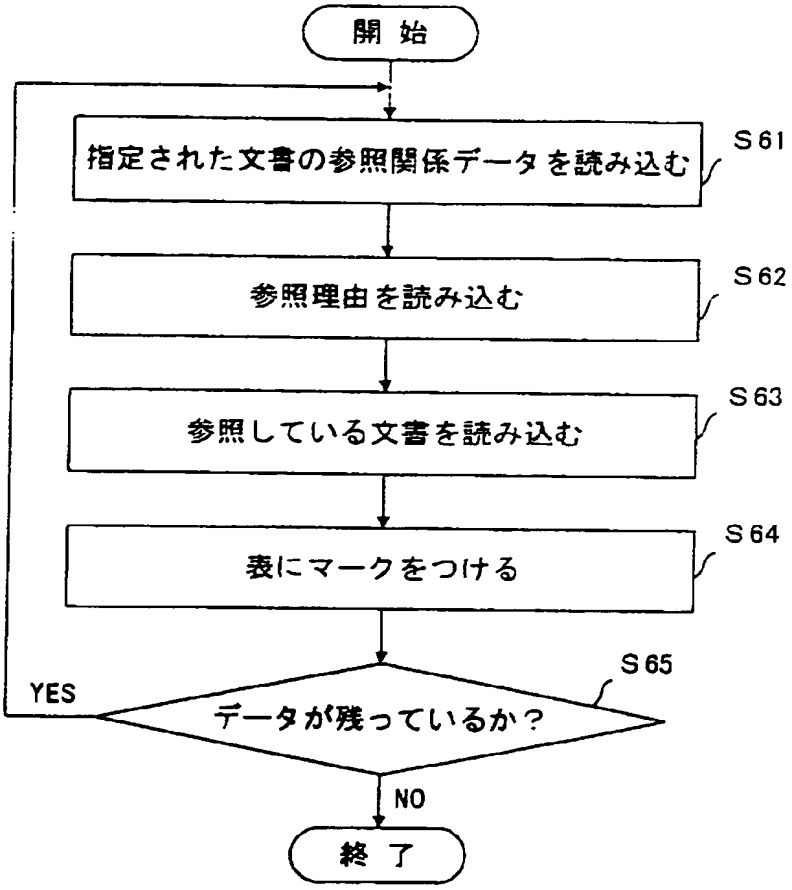
【図１２】

第１の検索結果表示を示す図



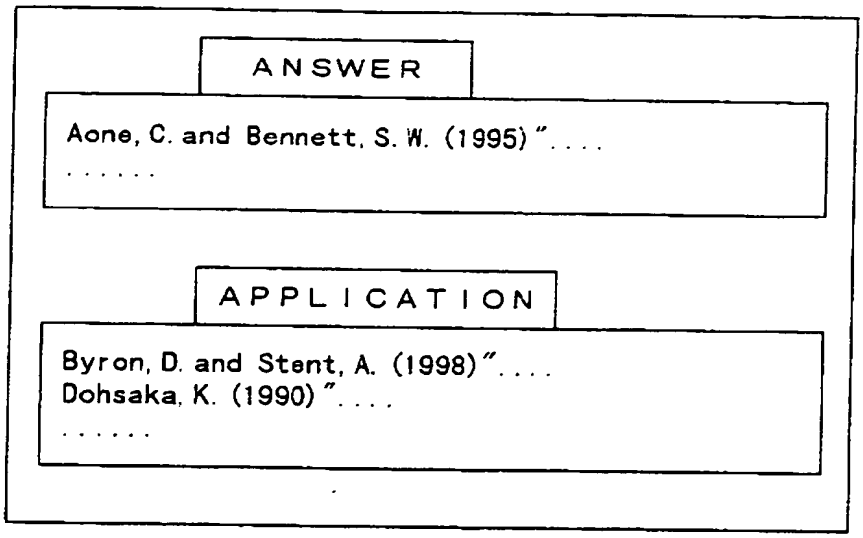
【図１３】

表示処理のフローチャート



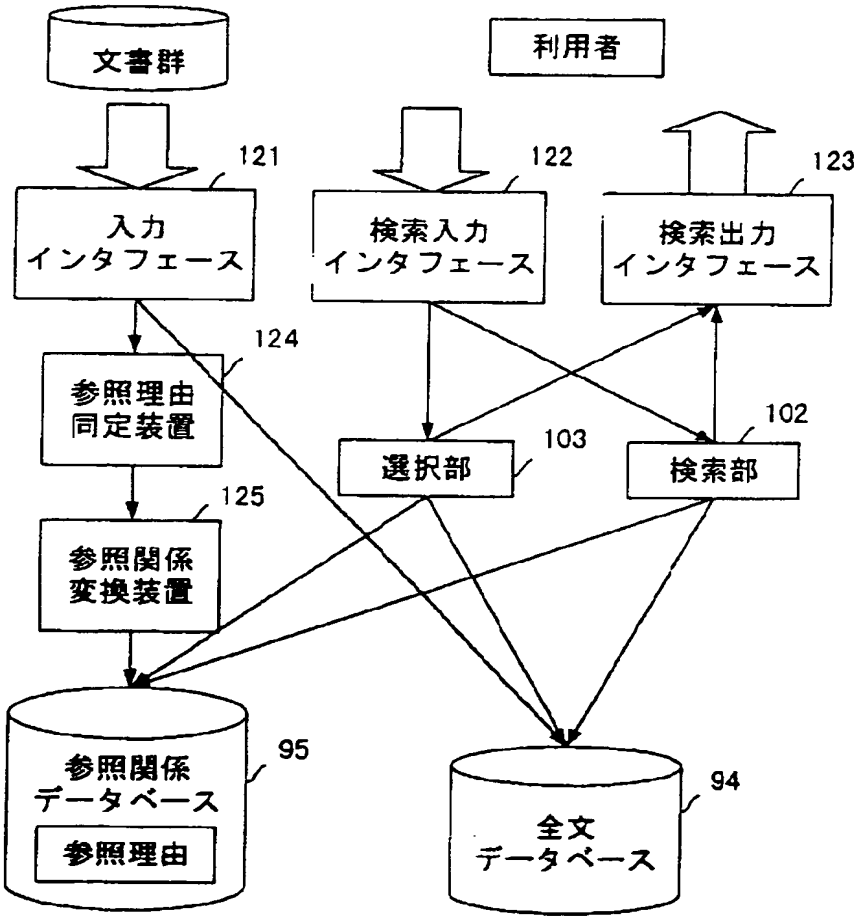
【図１４】

第２の検索結果表示を示す図



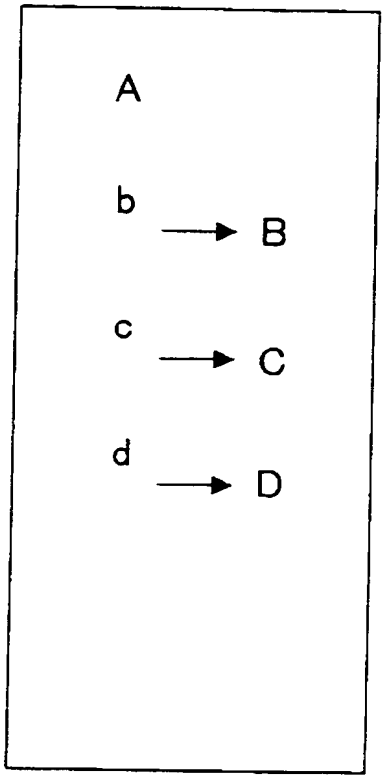
【図１５】

情　報　提　示　装　置　の　構　成　図



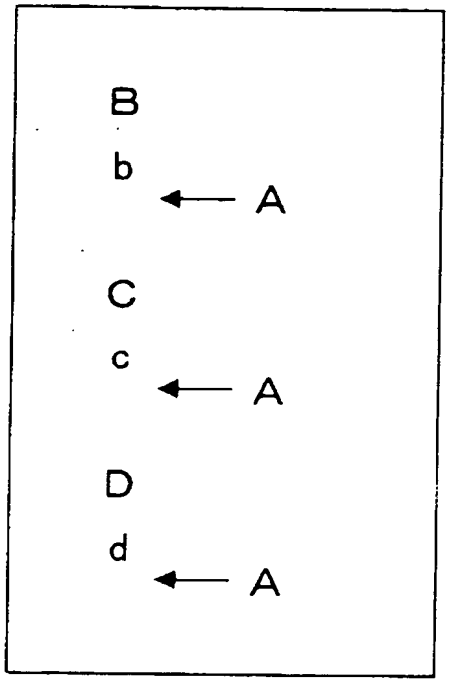
【図１６】

第3の参照関係を示す図



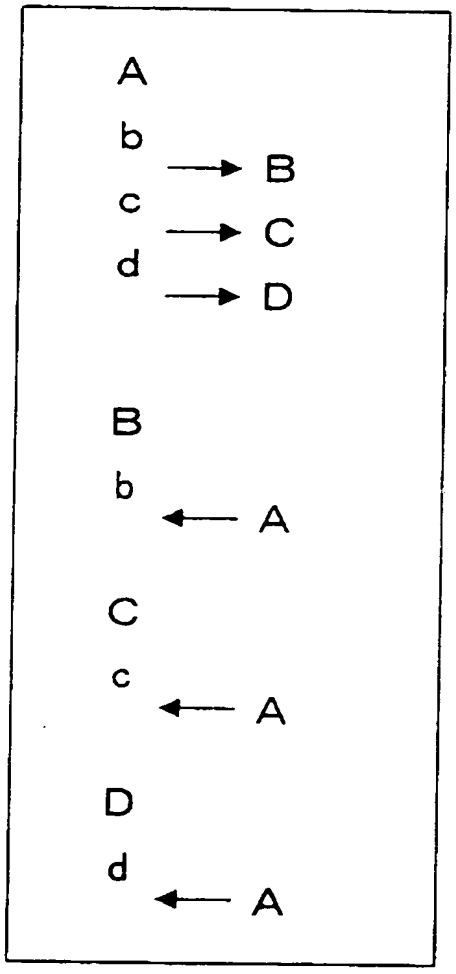
【図１７】

第4の参照関係を示す図



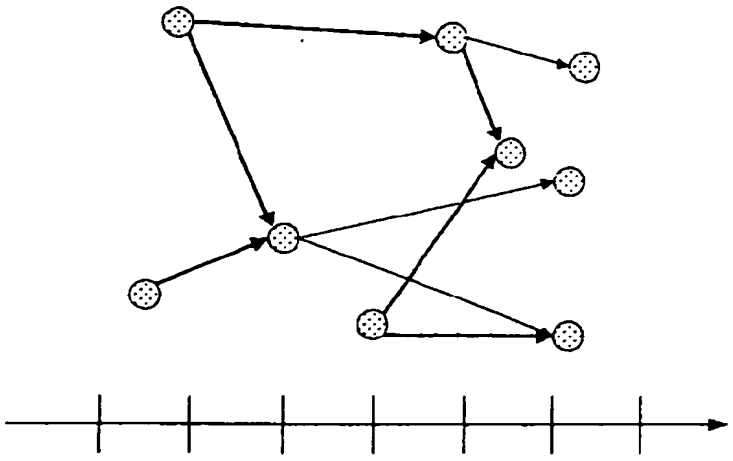
【図１８】

第5の参照関係を示す図



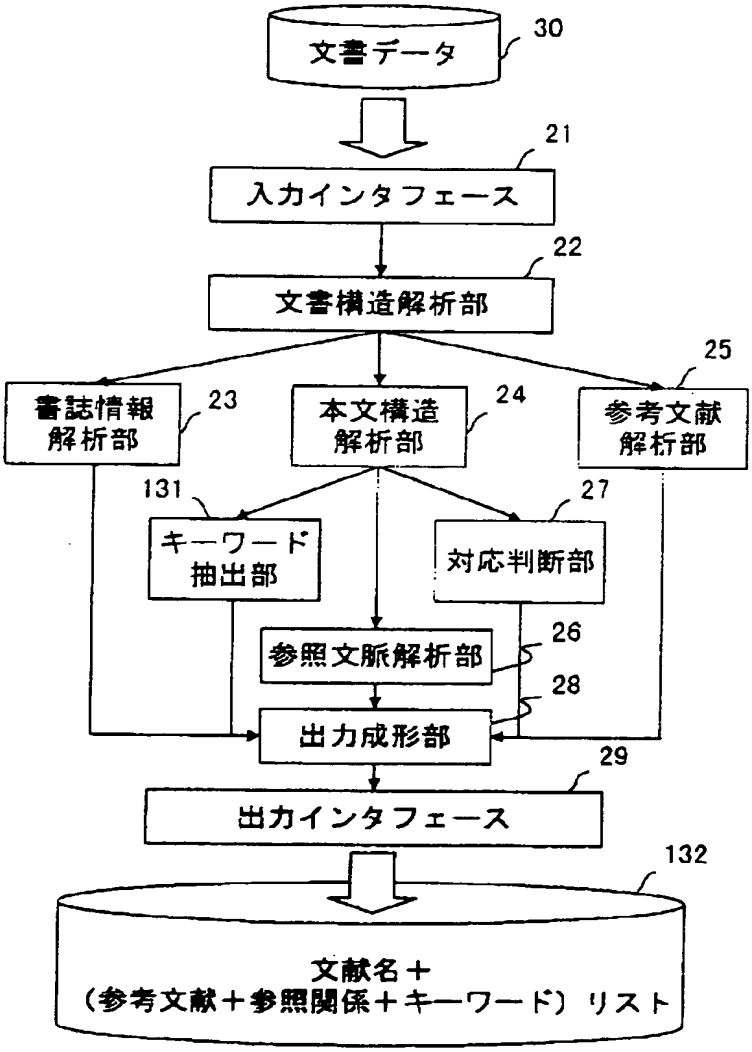
【図１９】

第１の時系列表示を示す図



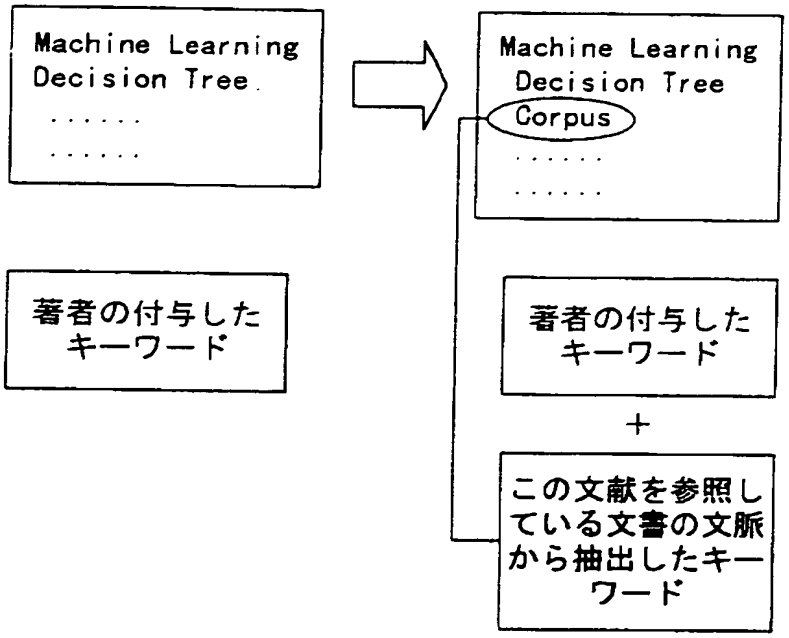
【図２０】

参照キーワード抽出装置の構成図



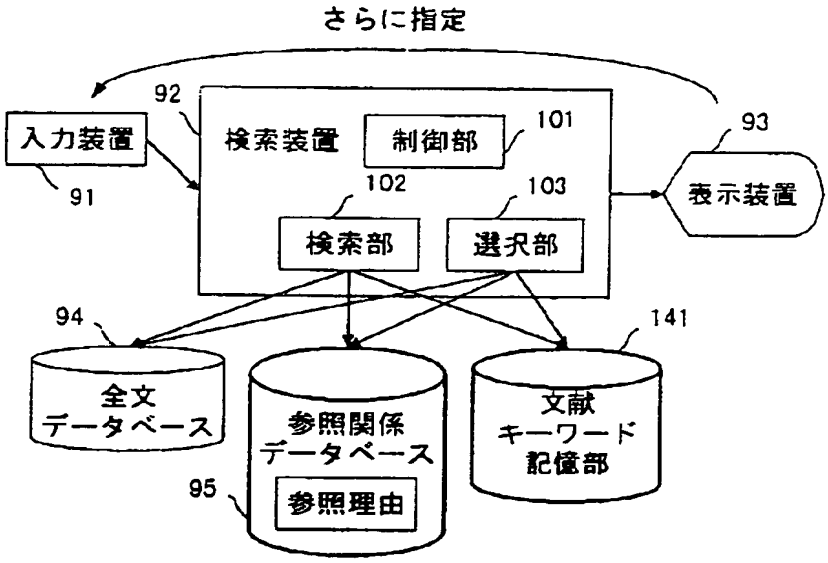
【図２１】

参照関係を用いたキーワードを示す図



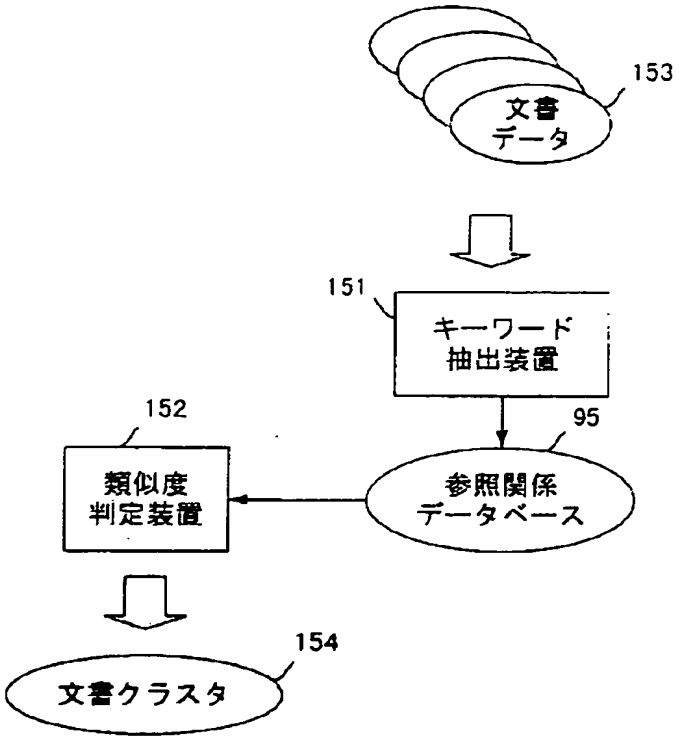
【図２２】

キーワードを用いた検索を示す図



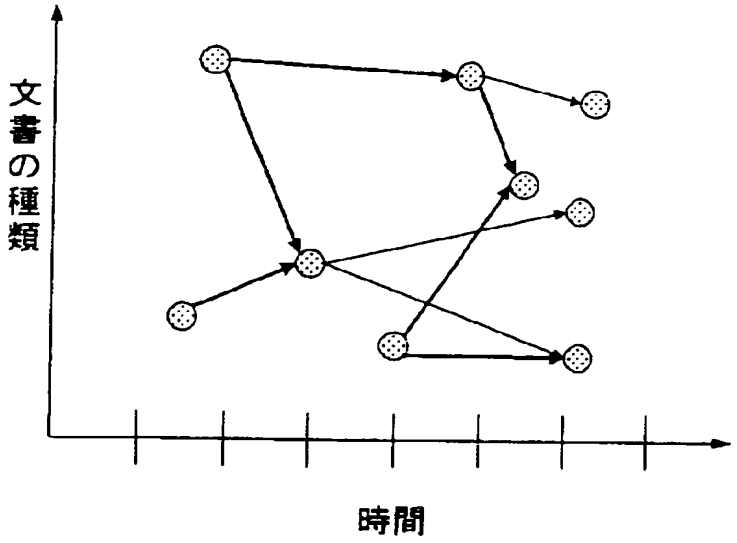
【図２３】

文書分類装置の構成図



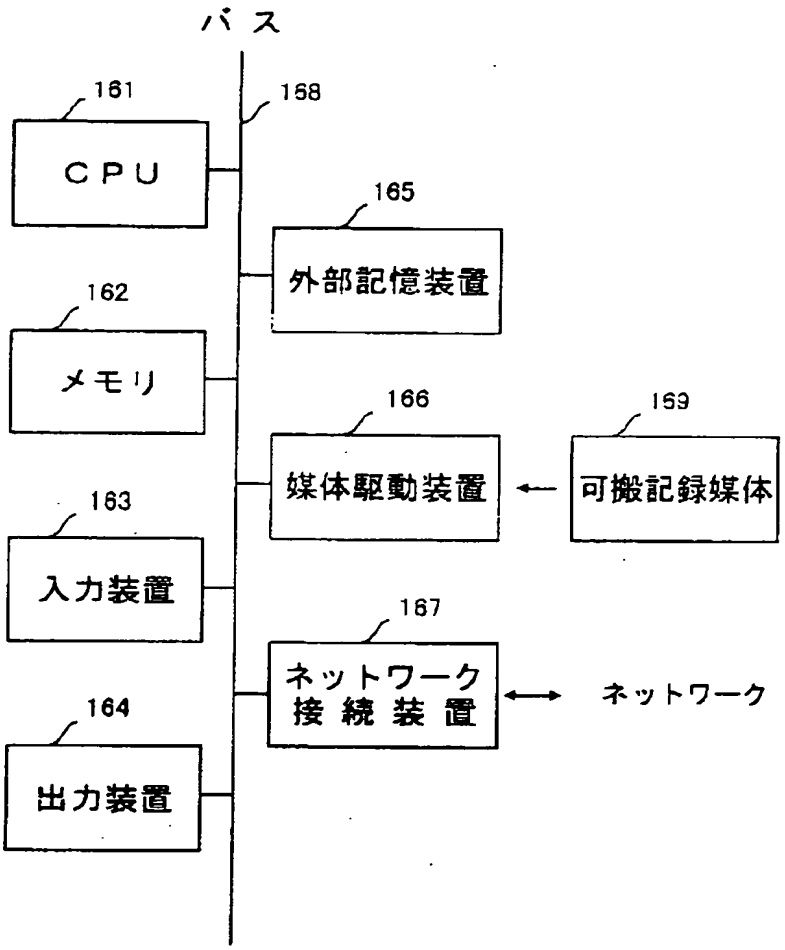
【図２４】

第２の時系列表示を示す図



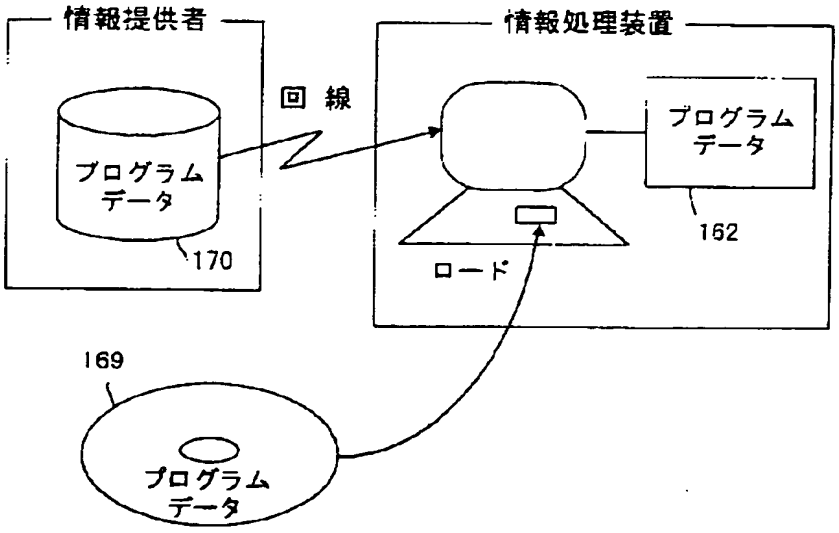
【図２５】

情　報　処　理　装　置　の　構　成　図



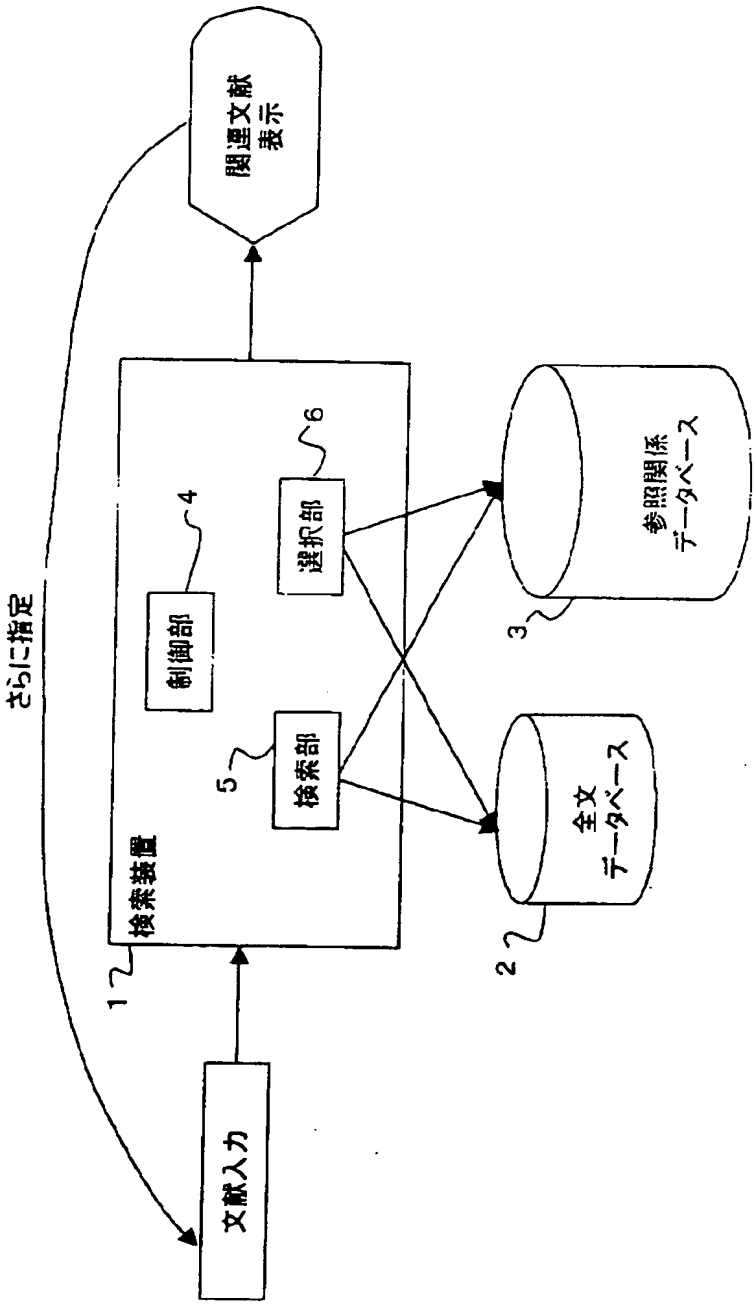
【図２６】

記　録　媒　体　を　示　す　図



【図２７】

従来の情報検索システムを示す図



【図２８】

従来の表示形式を示す図

重要文書（似たような参照、共引用）
Aone, C. and Bennett, S. W. (1995)".... Byron, D. and Stent, A. (1998)".... Dohsaka, K. (1990)"

【書類名】　要約書

【要約】

【課題】　ある文書が他の文書を参照している理由を同定し、同定された参照理由を利用して効率的に文書を検索することが課題である。

【解決手段】　文書構造解析部２２は、文書データ３０を書誌情報、本文、参考文献の３つの部分に分割し、それぞれ、書誌情報解析部２３、本文構造解析部２４、参考文献解析部２５に与える。参照文脈解析部２６は、本文構造解析部２４の解析結果から参考文献の参照理由を判断し、出力成形部２８は、参照理由を含む参照関係の情報を出力する。得られた参照関係を用いて、文書検索、文書分類等が行われる。

【選択図】　図２

出　願　人　履　歴　情　報

識別番号　　〔０００００５２２３〕

１．変更年月日　　１９９６年　３月２６日
　［変更理由］　住所変更
　　住　所　　神奈川県川崎市中原区上小田中４丁目１番１号
　　氏　名　　富士通株式会社